الجمهورية الجزائرية الديموقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

Département d'Électronique

End-of-study project dissertation

For the attainment of the State Engineer Degree in Electronics

Model Based Deep Learning for Computational Imaging: Application to Robust Multimodal 3D Imaging

Ouarda MEKERRI & Ilhem Meroua KACI
Under the supervision of :
Pr. Abderrahim Halimi & Pr. Mohamed Oussaid Taghi

Presented and defended publicly on 24/06/2025.

President:	Pr. Cherif LARBES	ENP
Supervisor:	Pr. Abderrahim HALIMI	HWU
Co-Supervisor:	Pr. Mohamed Oussaid TAGHI	ENP
Examinatrice:	Pr. Nesrine BOUADJENEK	ENP

ENP 2025
10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Algiers, Algeria.
www.enp.edu.dz

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

Département d'Électronique

End-of-study project dissertation

For the attainment of the State Engineer Degree in Electronics

Model Based Deep Learning for Computational Imaging: Application to Robust Multimodal 3D Imaging

Ouarda MEKERRI & Ilhem Meroua KACI
Under the supervision of:
Pr. Abderrahim Halimi & Pr. Mohamed Oussaid Taghi

Presented and defended publicly on 24/06/2025.

President:	Pr. Cherif LARBES	ENP
Supervisor:	Pr. Abderrahim HALIMI	HWU
Co-Supervisor:	Pr. Mohamed Oussaid TAGHI	ENP
Examinatrice:	Pr. Nesrine BOUADJENEK	ENP

ENP 2025
10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Algiers, Algeria.
www.enp.edu.dz

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
École Nationale Polytechnique

Département d'Électronique

Mémoire de fin d'études

En vue de l'obtention du diplôme d'ingénieur d'État en électronique

Apprentissage profond guidé par modèle pour l'imagerie computationnelle : Application à l'imagerie 3D multimodale robuste

Ouarda MEKERRI & Ilhem Meroua KACI
Sous la direction de :
Pr. Abderrahim Halimi & Pr. Mohamed Oussaid Taghi

Présenté et soutenu publiquement le 24/06/2025.

Président :	Pr. Cherif LARBES	ENP
Encadrant :	Pr. Abderrahim HALIMI	HWU
Co-encadrant :	Pr. Mohamed Oussaid TAGHI	ENP
Examinatrice :	Pr. Nesrine BOUADJENEK	ENP

ENP 2025
10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger, Algérie.
www.enp.edu.dz

ملخص

يعد التصوير ثلاثي الأبعاد أداة محورية لفهم وتحليل البيئات الواقعية بدقة. وتوفر تقنية LiDAR ، الذي يعتبر نظام لكشف وتحديد المدى باستخدام الضوء (الليزر)، حلاً متقدماً لإنتاج سحب نقطية دقيقة تُستخدم في مجالات متعددة. لكن هذه البيانات غالباً ما تتأثر بانخفاض الدقة والضجيج، مما يعيق استخدامها في بيئات معقدة. لمعالجة ذلك، يُقترح نهج يعتمد على خوارزميات التعلم العميق والمعالجة متعددة النطاقات بهدف تحسين جودة السحب النقطية. وقد أثبتت التجارب على بيانات محاكاة فعالية هذا النهج في تقليل الضجيج وتحسين التمثيل ثلاثي الأبعاد مع الحفاظ على التفاصيل الدقيقة.

الكلمات المفتاحية : التصوير ثلاثي الأبعاد، LiDAR ، السحب النقطية، التعلم العميق، المعالجة متعددة النطاقات.

Résumé

L'imagerie 3D occupe une place centrale dans de nombreuses applications où la précision spatiale est cruciale. Les capteurs LiDAR s'imposent comme une solution de référence grâce à leur fiabilité et leur capacité à mesurer des distances avec une grande exactitude. Toutefois, dans des conditions d'acquisition réalistes, leurs performances peuvent être compromises par la présence de bruit et la faible densité d'informations issues de mesures limitées.

Pour répondre à ces défis, nous avons développé une approche reposant sur l'apprentissage profond, combinée à un traitement multi-echelle, afin d'améliorer la qualité de reconstruction des images de profondeur. Cette méthode, évaluée sur des données LiDAR simulées, a démontré des gains significatifs en précision et en robustesse, même dans des contextes bruités.

Mots clés : Carte de profondeur, Nuage de points, Lidar monophoton, Dtof, Multi-échelle, Robuste, Rareté photonique.

Abstract

3D imaging is critical in applications requiring precise spatial detail. Among available technologies, LiDAR sensors are particularly prized for their accuracy and reliability. However, in realistic conditions, their performance is usually compromised by photon noise and sparse, low-resolution measurements.

To overcome these limitations, we introduce a deep learning approach with multiscale processing to produce high-quality depth reconstructions despite low-quality input data. Tested on simulated LiDAR datasets, the method has notable improvements in accuracy and robustness.

Keywords : Depth map, Point cloud, single photon Lidar, Dtof, Multi-scale, Robust, Photon sparsity.

Acknowledgments

We would like to extend our heartfelt and deepest gratitude to the best supervisor, Mr. Abderrahim HALIMI, Despite his many responsibilities, he generously dedicated his time to guide us throughout the project, standing by our side until the very end. His insightful advice, constructive feedback, and kindness were invaluable to our progress. The difference between where we were at the beginning of February and where we are now is a clear sign of great progress. Thank you so much, sir.

We would like to extend our appreciation to our co-supervisor, Mr.TAGHI Mohamed, for his continuous guidance and support throughout the entire journey of this thesis. His availability and advice played a crucial role in the successful completion of this work.

We extend our sincere thanks to the President of the jury, Mr. LARBES Cherif, and to the Examiner, Mrs. BOUADJENEK Nesrine, for accepting our work.

Ouarda & Meroua "KIMEK"

Dedication

Most importantly, Al-Hamdulillah, all glory to **god.** *What I have done would never be possible without him on my side. Every step, every blessing, every success, and every choice I made were all made possible thanks to* ***god,*** *he answered my prayers.*

To my beloved parents,

Words fall short when it comes to expressing the depth of my gratitude. Nothing in this journey or in my life would have been possible without your constant love, unwavering support, and infinite sacrifices. You stood by me in moments of doubt and discouragement, lifted me up when I was at my lowest, and celebrated with me in moments of joy and success. Your faith in me never wavered, even when mine did. You have given me the strength to persevere, the courage to dream, and the values to walk through life with dignity. This accomplishment is as much yours as it is mine the fruit of every sleepless night you endured for me, every word of encouragement you whispered, and every silent prayer you made.

Thank you for being my safe haven, my source of light, and my greatest blessing. I hope to one day make you as proud as I am grateful.

To my dear grandmothers — Mami and Omi *— thank you for your endless prayers and tenderness. I wish Mami could be present with us today, but I know her douas continue to surround me with strength.*

To my late grandfather, Baba Sido *— may Allah grant him mercy — your memory and wisdom remain a guiding light in my life.*

To my entire family: *To all my uncles, aunts, and cousins, thank you without exception for your constant encouragement, kind words, and the love you've shown throughout this journey. Every message, every visit, every prayer meant more than you know.*

A heartfelt thank you to ***my cousins: Islam, Lokmane , Mohamed, Salim and Ishak*** *who have always stood by me not only as family but as true siblings sharing laughs, lifting my spirits, and offering unwavering support whenever I needed it. Your presence in my life has been a gift, a reminder that no matter the distance or time, family remains our greatest strength.*

To my dearest friends,

Fella, *even though we never sat together in a single classroom, we've shared something far more meaningful — life itself. You've been a constant presence, a comforting soul, and I couldn't be more grateful for your friendship.*

Soundous, *one of the most precious gifts that ENP has given me. Your smile has brightened countless days, and your presence has turned simple moments into lasting memories. I truly cannot imagine this journey without you by my side.*

Sonia, *I miss our endless laughter and those inside jokes only we could understand. I wish you could be here with me to share this milestone but your presence is always felt, no matter the distance. I hope we'll see each other again soon, and pick up right where we left off.*

Your friendships mean the world to me. I don't have sisters by blood, but God gave me something just as special you.

To all my friends at ENP, especially Leila, and to everyone from the Electronics Department with a special mention to Houda and Wissem thank you for the laughter, the memories, and every moment that made these years so meaningful. To the incredible class of 2022–2025, you've been one of the most beautiful parts of this adventure. I'll always treasure the friendships we built and the unforgettable experiences we shared.

To my partner *throughout this incredible journey,* ***Ouardaty*** *I feel truly fortunate to have walked this path alongside you. You've been more than just a colleague; you've been a constant source of strength, motivation, and calm in every moment of doubt or pressure. Your presence turned challenges into shared victories, and long days into unforgettable memories. From our first brainstorming sessions to the final steps of this project, thank you for your patience, and your friendship. This achievement is as much yours as it is mine.*

KIM

Dedication

لا يزال صدى الخطوات المتسارعة نحو باب الابتدائية محفورًا في الذاكرة، حين اصطحبني أبي وسط دعوات أمي لي، وكأنّ ذلك كان بالأمس، ولم تمضِ عليه سبع عشرة سنة من السير المتواصل. كان السير على هذا الدرب مليئًا بالإنجازات وتعثرات الإخفاق، بكثير من الصبر والصمود والصراع. سبع عشرة سنة مضت على هذا الطريق، بعون الله وفضله وتوفيقه، فالحمد لله الذي يسّر لنا سبل النجاح.

إلى والديَّ... سندي، وقوتي، وإلهامي، ومصدر إرادتي. لا حروف تعبّر عن صنيعكما لي، ولا كلمات تفي بوصف جميلكما، سأصمت لعل السكوت يحتضن الامتنان الذي عجز لساني عن قوله.

إلى إخوتي هاجر و عبد السلام، ورغم كثرة الشجارات، إلا أن فضلكما عليّ لا يُقدّر بأغلى الأثمان، فشكرٌ موصول لكما على تحمّلكما مزاجي وتقلّب نفسي، وإنّي لأعلم أنّ الشكر وحده لا يكفي لردّ المعروف.

الى صديقتي حياة شكرا على المواقف الجميلة التي تشاركناها، تبادلنا للهموم على ضفاف شاطئ البحر و دردشتنا هناك لساعات لقد كانت لحظات جميلة حقا.

صديقتي سامية، رغم قلة لقاءاتنا الا انك ستضلين صديقة طفولتي الغالية شكرا لوجودك صديقتي.

إلى زميلاتي وأحلى صدف دربي: وسام، مروة، هدى، وصونيا التي غادرتنا باكرًا... تشاركنا القطار نفسه، ضحكنا وبكينا، وتقاسمنا الهموم ومواقف الحياة. لقد كانت سنين لا تُنسى.

إلى زميلتي وصديقتي في هذا المشروع، مرويتا، أتذكرين ذلك اليوم الذي جلسنا فيه بمحاذاة الطريق، نغالب الغضب والحسرة، حائرَتَين فيما سنختاره؟ مضت ستة أشهر منذ ذلك الحين، وها نحن اليوم نطوي صفحة، لنفتح أخرى جديدة. بيننا الكثير من اللحظات والمواقف التي لا يدركها سوانا، نحن فقط نعلم تمامًا ما مررنا به، تجاوزنا الكثير أما القادم فأجمل.

ودون أن أنسى... My best computer, كنتَ صديقًا شهد كل حالاتي، فشكرًا لك.

و بالطبع وردة بنسختيها الجديدة و القديمة شكراااا لك.

MEK

Contents

List of Tables

List of Figures

List of Acronyms

- **LIDAR :** Light Detection and Ranging
- **CI :** Computational Imaging.
- **MBDL :** Model-Based Deep Learning.
- **DVSR :** Depth video super resolution.
- **HVSR :** Histogram video super resolution.
- **ToF :** T-of-flight.
- **DTOF :** Direct time of flight.
- **RMSE :** Root mean square error.
- **AE :** Absolute error.
- **MF :** Matched filter.
- **MAE :** Mean Absolute Error.
- **ARGMAX :** argument of the maximum.
- **PPP :** Photons per-pixel.
- **SBR :** Signal-to-Background Ratio.
- **UC-Ne :** Uncertainty inspired RGB-D saliency detection.
- **RGB :** Red,green and blue.
- **RGB cameras :** Red-Green-Blue cameras.
- **AR :** Augmented reality.
- **VR :** Virtual reality.
- **SPL :**Single-Photon LiDAR.
- **TCSPC :**Time-Correlated Single-Photon Counting.
- **SPAD :** Single-Photon Avalanche Diode.
- **CMOS :** Metal-oxide-semiconductor.
- **GPS :** Global Positioning System.

- **IMU :** Inertial Measurement Unit.
- **GPU :** Graphics Processing Unit.
- **MRI :**Magnetic Resonance Imaging.
- **AI :** Artificial Intelligence.
- **FMCW :** Frequency Modulated Continuous wave.
- **AMCW :** a Amplitude-Modulated Continuous Wav.
- **SNR :** signal-to-noise ratio.
- **Replica :** Realistic perception for learning indoor and complex Activities.
- **ARKit :** Apple's Augmented Reality Kit.
- **DyDToF :** Dynamic Direct time of flight.
- **3D :** Three-dimensional.
- **2D :** Two-dimensional.
- **DCT :** Discrete Cosine Transform.
- **PCA :** Principal Component Analysis.
- **CNNs :**Convolutional Neural Networks.
- **VGG :**
- **PSNR :** Peak Signal to Noise Ratio.
- **SSIM :** structural similarity index measure.
- **MSE :** Mean squared error.
- **GANs :** Generative ad- versarial networks.
- **NLP :** Natural language processing.
- **PU-Net :** Point Cloud Upsampling Network.
- **EMD :** Earth Mover's Distance.
- **NUC :** Normalized Uniformity Coefficient.
- **PU-GCN :** Point Cloud Upsampling using Graph Convolutional Networks.
- **GCNs :** Graph Convolutional Net-works.
- **MLP¨:** Multi-Layer Perceptron.
- **PU1K :** Point Cloud Upsampling 1000 Dataset.
- **PU-GAN :** Point Cloud Upsampling Generative Adversarial Network.
- **3PU :** Progressive Point Cloud Upsampling.
- **CVAEs :** conditional variational autoencoders.
- **TEPE :** Temporal end-point error.

- **HD :** Hausdorff distance.
- **CD :** Chamfer distance.
- **P2F :** Point-to-surface distance.
- **SSIM:** Structural Similarity.

General Introduction

In today's fast-evolving technological landscape, the way we perceive and interpret the world around us is being fundamentally reshaped by advances in **computational imaging**. Whether it's enabling self-driving cars to safely navigate busy streets, helping robots understand and interact with their surroundings, or supporting immersive augmented reality experiences computational imaging lies at the heart of it all. What makes it so powerful is its unique ability to go beyond the limitations of traditional cameras and sensors, by combining **physical models** with **intelligent algorithms** to reconstruct, enhance, and analyze visual information in ways that were once considered impossible.

One of the most exciting and impactful areas within this field is **3D imaging** the ability to capture not just reflectivity images, but depth and structure, giving machines a true sense of space. At the core of many 3D imaging systems is **LiDAR (Light Detection and Ranging)**, a technology that uses laser pulses to measure distances and generate precise depth maps. LiDAR has already proven invaluable in a wide range of fields, from autonomous vehicles and drones to smart cities and environmental monitoring. However, like any technology, it comes with its challenges. LiDAR sensors can be expensive, they often suffer from noise and limited resolution, and in many real-world scenarios, they simply can't provide the full picture on their own.

To overcome these limitations, researchers and engineers are increasingly turning to **multimodal imaging** combining LiDAR data with other sources like RGB images, thermal data, or radar signals—to create richer, more reliable representations of the environment. But merging and making sense of these different data types isn't easy. It requires advanced methods that are not only data-driven, but also guided by physical understanding. This is where **Model-Based Deep Learning (MBDL)** enters the scene.

MBDL is a hybrid approach that blends the strengths of two worlds: the **rigor and interpretability** of traditional model-based methods, and the **learning power and adaptability** of deep neural networks. Rather than treating deep learning as a black box, MBDL embeds known physics and mathematical constraints into the learning process, leading to smarter, more robust models that can work well even in challenging conditions like noisy data, missing information, or limited training examples.

This project explores how MBDL can be applied to improve robust **multimodal 3D imaging**, especially by leveraging the strengths of LiDAR alongside complementary data. Our goal is to design a practical, theoretically sound framework that makes 3D imaging more accurate, more reliable, and more adaptable to real-world conditions. Throughout this work, we'll delve into the theory behind these methods, explore existing solutions and their limitations, and propose a novel approach backed by experimental validation.

Ultimately, this study aims to contribute not just to academic knowledge, but also to the development of intelligent imaging systems that can truly understand the complex world we live in.

To explore this promising direction, the present work is structured across five chapters :

It begins with a general introduction outlining the field and its main motivations.

Chapter 1 This chapter delineates the foundational context and articulates the scientific rationale for integrating deep learning with physics-based modeling as a means to address current limitations in 3D perception.

Chapter 2 presents a detailed analysis of LiDAR system architecture, data quality constraints, and the influence of noise and signal parameters on depth reconstruction accuracy.

Chapter 3 reviews existing methods both traditional and deep learning-based—and high- lights the promise of hybrid approaches that merge the strengths of both worlds.

Chapter 4 outlines our proposed method, including the problem setup, the model archi- tecture, and strategies to improve robustness and adaptability.

Chapter 5 analyzes the results, discussing the model's performance, its strengths, and areas for improvement.

Finally, a general conclusion summarizes the study and outlines future research directions.

CHAPTER 1 :

Context and challenges in computational 3D vision

In a world where vision is quickly becoming the cornerstone of most emerging technologies, computational imaging has been developed as a cross-disciplinary science that combines physics, signal processing, and machine learning. Unlike classical approaches, it applies cutting-edge algorithms to extract rich information from raw data and thereby becomes highly suitable for applications such as robotics, autonomous driving, smart surveillance, and medical imaging. This chapter establishes the general context of multimodal 3D vision, emphasizing the complementarity of LiDAR, RGB cameras, and radar sensors, and presents our hybrid approach we explain, based on model-based deep learning with physical principles incorporated into learning. It also defines the key problem addressed in this study, outlining the limitations of current methods and the specific aims aimed at reinforcing environmental perception in complex and dynamic contexts.

1.1 Context and challenges of LiDAR-based 3D imaging

Despite the development of 3D sensing technologies, LiDAR systems, particularly Direct Time-of-Flight (DTOF)-based ones, are still challenged by low spatial and temporal resolution. Since there are trade-offs between acquisition speed and power consumption versus accuracy, point clouds acquired tend to be sparse and noisy, negatively impacting the quality of 3D scene understanding, and small or distant object detection and accurate segmentation become challenging.

Improved LiDAR resolution is especially important in real-time and long-range applications, where accurate geometric information matters. However, whereas improved data quality can be delivered through enhanced sensor hardware, this incurs a cost in terms of practicability in most cases. Therefore, the inherent challenge is to design computational methods able to derive high-resolution 3D data from low-quality measurements.

This work belongs to the broader class of computational imaging, seeking to extract rich information from unprocessed data based on algorithmic modeling. Instead of relying solely on physical sensor innovation, we highlight learning-based approaches to recover subtle 3D details.

Our goal is to enhance LiDAR depth data resolution and quality by combining it with high-resolution RGB data. Through a model-driven deep learning architecture, we leverage both the physical properties of the sensing process and the learnable aspects of neural networks. This combination enables super-resolution, denoising, and improved depth estimation from sparse data.

Lastly, our goal is to bridge the gap between low-cost, low-resolution sensing and high-fidelity 3D perception—providing more robust, higher resolution insight into complex scenes.

1.2 Motivation and approach

The proposed method combines cutting-edge multimodal sensor fusion with Model-Based Deep Learning (MBDL) framework to address growing perception task complexity in challenging dynamic situations. By fusing LiDAR, RGB camera, and radar data, the method leverages each modality's strengths: LiDAR offers dense point clouds in 3D; RGB cameras offer rich color and texture for object detection; and radar provides strong motion detection despite unfavorable weather conditions.

Historical perception methods based only on physical or statistical models struggle to generalize across real-world variability, while completely data-driven deep learning techniques have been shown to need large amounts of data and lack interpretability. Our approach, however, takes the MBDL paradigm that integrates physical and mathematical priors within deep learning models. This integration not just regularizes and directs the learning process but also enhances generalization, robustness, and data efficiency.

In order to attain optimal sensor complementarity, we rely on sophisticated neural architectures capable of processing varied streams of data. These models combine geometric, visual, and dynamic information for improved object detection, segmentation, and 3D scene interpretation. Optimized alignment algorithms, synchronization, and real-time fusion facilitate the combination process, resulting in consistent performance even under adverse environments such as fog, rain, or sensor misalignment.

Second, our approach utilizes a modular Plug-and-Play architecture that treats each processing step—depth refinement, saliency detection, or anomaly processing—as an independent but compatible module. Modularity enables plugin-compatible model-driven and data-driven modules so that the system can react in an adaptive way to environmental changes and sensor faults. The result is an efficient, explainable, and fault-tolerant perception pipeline for challenging applications in autonomous driving, robotics, and intelligent surveillance.

1.3 Project objectives

This study sets out to achieve several specific objectives aimed at advancing the field of multimodal 3D imaging using model-based deep learning techniques.

- The project consists of developing advanced computation algorithms that combine the rigorousness of statistical models with the learning capacity of deep neural networks. The goal is to create a hybrid system that attempts to solve complex inverse problems related to 3D imaging.
- Our work aims to enhance spatial resolution under sparse-photon conditions and strong background noise (e.g., fog or rain), enabling accurate long-range imaging at high frame rates.
- The study also emphasizes the fusion of heterogeneous sensor data—integrating information from single-photon LiDAR and RGB cameras. This multimodal fusion is expected to significantly reduce uncertainties and improve reconstruction fidelity.
- Finally, the work also studies high-level computer vision tasks such as saliency detection on RGB-D data, object detection, and segmentation. These tasks make use of high 3D reconstructions for the purpose of improving scene understanding and allowing generalization to real-world tasks such as robotics, augmented reality, smart surveillance, and human–computer interaction.

1.4 Conclusion

Finally, chapter one established the theoretical and technical foundation of the project by introducing the issues of modern 3D imaging and the limitations of traditional methods. It emphasized the significance of hybrid approaches combining physical modeling with deep learning, as well as the advantages of multimodal sensor fusion in enhancing perception systems. These served as the context for chapters two to four, which will detail the approach and deliver results in real applications.

CHAPTER 2 :

Background Knowledge

2.1 Introduction

The evolution of LiDAR systems has transformed 3D perception on the basis of advanced opto-electronic components and increasingly powerful computational imagery. Each subunit ranging from the laser source to GPS/IMU modules is tasked with producing accurate point clouds. With advanced processing algorithms, these devices enable not just the capture of dense spatial data but also depth image reconstruction via accurate computational models. This chapter discusses the major elements of a LiDAR system and how they play a role in creating three-dimensional computational images, which open doors to applications in mapping, autonomous navigation, and computer vision.

2.2 Computational imaging

2.2.1 Definition of computational imaging:

Computational Imaging (CI) is an advanced imaging approach that combines optics, sensors, and computational processing to overcome the limitations of traditional imaging systems. Unlike conventional methods that passively record light using sensors, CI actively modifies and encodes light information before capture and then applies sophisticated reconstruction algorithms to extract more detailed and useful images.

By integrating physical models, optimization techniques, and artificial intelligence algorithms, CI reconstructs images with unprecedented accuracy, even from imperfect or incomplete measurements. The rise of multi-core processors and GPUs has significantly enhanced the feasibility of CI, enabling real-time or near-real-time image reconstruction and enhancement [1].

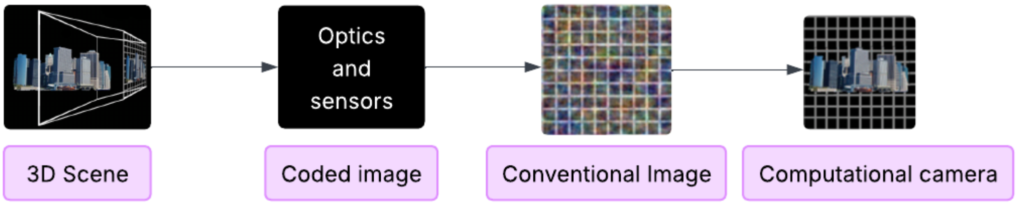


Figure 2.1: Principle of Computational Imaging

This Figure 2.1 illustrates the process of image acquisition and reconstruction using a computational camera system. It highlights the transformation of a 3D scene into a final conventional image through several key steps:

1. **3D Scene:** The real-world environment that is being captured.
2. **Perspective Projection:** The transformation of the 3D scene into a 2D representation based on geometric projection principles.
3. **Computational Camera:** A system that includes advanced optics and sensors designed to manipulate and encode visual information before capture.
4. **Coded Image:** The intermediate result obtained after the computational camera processes the light and scene information. This image is typically encoded in a way that allows for enhanced data extraction and post-processing.
5. **Post-processing:** Algorithms are applied to decode and reconstruct the final conventional image, restoring a viewable representation of the original scene.

This approach allows for improved image capture capabilities, such as enhanced resolution, depth perception, and noise reduction, making it particularly useful for applications in computer vision, robotics, and scientific imaging.

2.2.2 Applications of computational imaging :

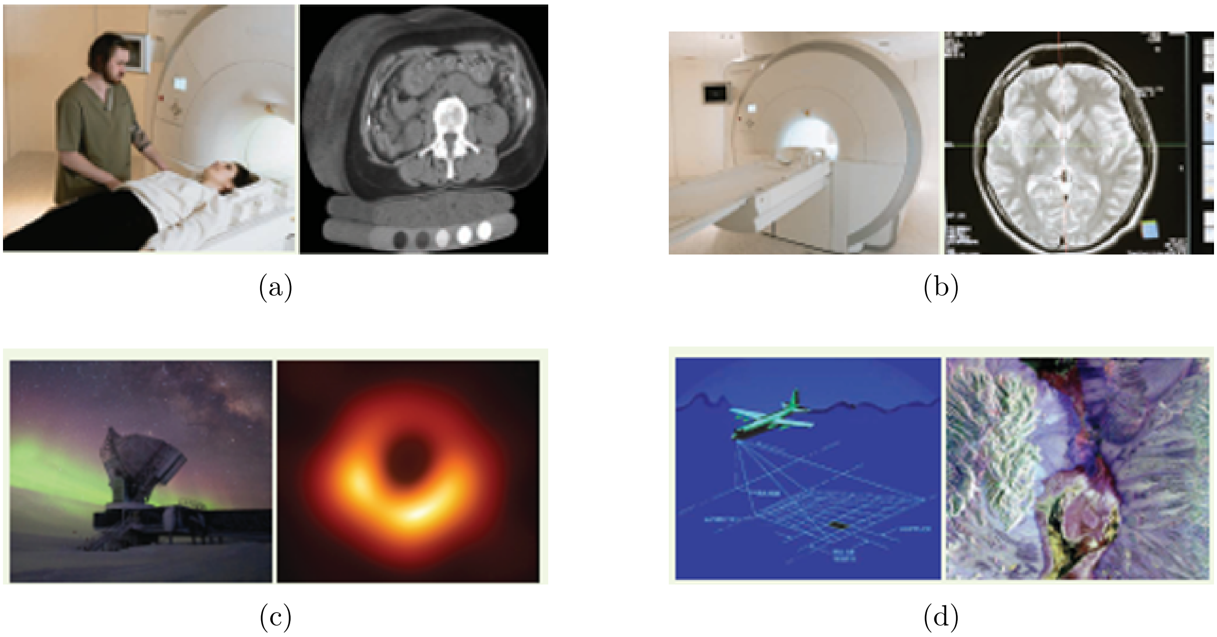
Computational Imaging (CI) is subtly transforming the way we observe and interpret the world around us.. As shown in Figures 2.2a and 2.2b, CI plays a critical role in the medical field by significantly improving the accuracy and quality of diagnostic imaging such as MRI and ultrasound. Enhanced resolution in CT scans, MRIs, and microscopic images enables earlier disease detection and better patient monitoring.

At a much smaller scale, in advanced microscopy, CI makes it possible to visualize ultra-fine structures such as viruses and proteins—structures that were once invisible to conventional imaging methods—thus opening new frontiers in biology and medicine. In astronomy, CI pushes telescopes beyond their physical limitations, enabling groundbreaking milestones such as the first-ever image of a black hole (see Figure 2.2c). Likewise, radar imaging systems (Figure 2.2d) are enhanced by CI, improving detection capabilities in both scientific and industrial applications.

In industrial manufacturing, CI is key to quality assurance, enabling automated systems to detect even the tiniest product defects, thus elevating production standards. Environmentally, CI boosts satellite and remote sensing technologies, facilitating more effective ecosystem monitoring, smarter agriculture, and more precise geographic and climate mapping.

In everyday life, CI strengthens advanced security and surveillance systems—making facial recognition and object detection more robust, even in difficult lighting or weather conditions. In consumer photography, CI allows for post-capture adjustments like refocusing and dynamic lighting changes, enabling users to take stunning images in low-light settings.

In summary, Computational Imaging is not just a technological innovation—it's a powerful, evolving tool that bridges science, industry, and daily life. By making the invisible visible and the complex understandable, CI is transforming the way we explore and interact with the world [2].



(a)

(b)

(c)

(d)

Figure 2.2: Examples of imaging systems: (a) MRI scan image, (b) Ultrasound scan image, (c) Astronomical telescope image, (d) Radar imaging system [2]

2.2.3 Importance of computational imaging for 3D Imaging

Computational Imaging (CI) is reshaping the world of 3D imaging by breaking through the limitations of traditional methods. Conventional systems like stereo vision, structured light, and time-of-flight sensors often struggle with noise, occlusions, and environmental factors that can distort data. But with CI, these challenges are met head-on by blending advanced optical techniques with algorithmic reconstruction and deep learning models. This combination allows for the extraction and enhancement of 3D information in ways that were previously impossible. One of CI's standout strengths is its ability to improve depth accuracy and spatial resolution. By using methods like multi-view fusion, compressive sensing, and neural network-based depth estimation, CI can produce incredibly detailed 3D models, even in low light or when sensor data is sparse. Techniques such as wavefront coding and phase retrieval also help capture richer optical data, enhancing depth perception and producing clearer 3D representations.

Another area where CI excels is in tackling occlusions and noise—persistent problems in fields like medical imaging, robotics, and autonomous vehicles. Traditional systems often fail when parts of a scene are blocked or the environment is noisy, but CI overcomes these issues using advanced techniques like model-based reconstruction and sensor fusion. By combining data from LiDAR, hyperspectral sensors, and depth-aware neural networks, CI enables systems to fill in gaps and accurately reconstruct 3D scenes, even when parts of the environment are obstructed or the data is imperfect.

In the realm of real-time 3D imaging, CI has made huge strides, thanks to advances in AI and computational power. Technologies like GPU acceleration and neural radiance fields (NeRF) are allowing for lightning-fast 3D imaging, which is crucial for applications like augmented reality (AR), virtual reality (VR), and medical diagnostics. These advancements make it possible to generate high-quality, real-time 3D images with minimal latency, making CI incredibly valuable for dynamic environments where quick, accurate depth information is essential.

CI also shines in multi-modal 3D imaging, where it combines different types of sensors to provide deeper insights across various fields. For instance, in medical imaging, AI-powered 3D reconstructions from CT scans, MRIs, and optical coherence tomography (OCT) help doctors diagnose more accurately and plan surgeries more effectively. In robotics and autonomous systems, enhanced 3D perception allows robots to understand their surroundings better and navigate complex environments with ease. And in scientific and industrial fields, CI's role in 3D imaging is crucial—whether it's helping astronomers study black holes, aiding in nanoscale imaging for research, or detecting defects in manufacturing.

Looking to the future, CI is expanding the possibilities of 3D imaging in ways we could only dream of a few years ago. By combining data-driven AI with traditional physics-based models, CI is pushing the boundaries with innovations like holography and plenoptic cameras. These breakthroughs are laying the foundation for next-gen imaging systems that will offer even more precision, adaptability, and scalability, revolutionizing industries from healthcare to manufacturing and beyond [1].

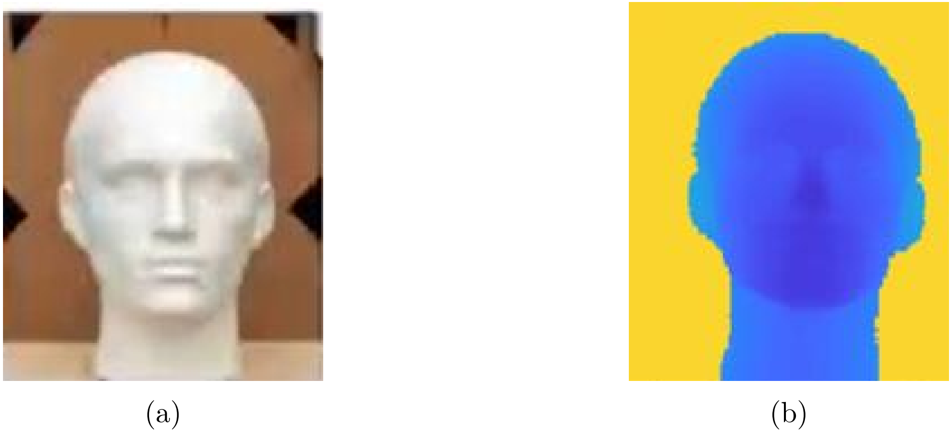
2.3 Fundamentals of Depth Maps and their applications

2.3.1 Definition

A depth map is a two-dimensional image where each pixel encodes the distance from a specific point in the scene to a reference viewpoint, typically the camera lens. This spatial information allows the depth map to represent the three-dimensional structure of a scene from a particular perspective. In computer vision and computer graphics, depth maps represent a fundamental intermediate representation that bridges the gap between 2D images and their underlying 3D geometries. When fused with RGB images, they allow for photorealistic 3D reconstructions, spatial comprehension, and interactive scene manipulation of real or virtual scenes. Depth maps can be created in a number of ways: through direct capture via 3D sensors like LiDAR, stereo cameras, or time-of-flight (ToF) sensors; through synthetic generation in simulators or 3D engines like Unreal Engine; or through multi-view reconstruction, in which depth is approximated by triangulation techniques analyzing multiple images captured from varying viewpoints.

2.3.2 The working principle of depth maps:

Every pixel in a depth map carries a numerical value that depicts the depth distance between the camera (or sensor) and the respective point in the scene. The value can be encoded as an 8-bit grayscale value, with pixel intensity ranging from black (close) to white (distant), or more accurately as a floating-point value that encodes the depth directly using physical units (e.g., meters or millimeters).Practically, the depth value at every pixel enables the system to calculate how distant the particular point is in 3D space from the camera position. Lower (darker) values generally designate closer objects, and higher (lighter) values designate more distant surfaces. The gradient of the depth value creates a depth-aware scene projection.



(a)

(b)

Figure 2.3: Example illustrating an RGB image (a) and its associated depth map (b), where blue indicates nearby objects and yellow indicates distant ones [3].

The illustration in Figure 2.3 illustrates how an ordinary RGB image 2.3a compares to its depth map. The colors in the depth map (b) 2.3b represent relative distances from the camera: blue hues are closer regions, green is for the middle distance, and red is for regions that are farther. The color gradient provides an intuitive and more informative visualization of spatial depth, and one can readily discern the 3D structure of the scene.

2.3.3 Applications of Depth Maps in vision and graphics

Depth maps are widely used in both computer vision and computer graphics for a variety of tasks. When combined with their corresponding RGB images, they enable the reconstruction of accurate 3D models of environments by assigning a depth value to each pixel, effectively transforming a flat image into a spatially coherent point cloud or 3D mesh. This representation is beneficial to applications like object tracking, augmented reality, collision detection, and robot navigation. Depth maps also facilitate high-level vision applications like object detection, pose estimation, and scene segmentation. In computer graphics, they facilitate photorealistic rendering effects like depth of field simulation, shadow mapping, subsurface scattering, and simulation of semi-transparent media like smoke or fog. They are also responsible for maximizing rendering efficiency through z-buffering and z-culling, and are vital to the creation of 3D illusions in stereoscopy and autostereograms.

2.4 3D LiDAR imaging

3D LiDAR (Light Detection and Ranging) is an advanced active remote sensing technology that enables precise three-dimensional measurements by analyzing the time of flight of laser pulses reflected off objects and surfaces. It employs laser scanners, also known as **laser radar, laser rangefinders, or laser profilers,** to capture high-resolution structural data from natural and urban environments.

Using measurement techniques such as **Time of Flight (ToF),** LiDAR can map terrains, forests, infrastructure, or moving objects with an accuracy ranging from **centimeters** to even **millimeters** in high-performance systems. Several categories of LiDAR exist:

- **Airborne LiDAR:** Initially developed for bathymetry and topographic mapping, this type is used for large-scale surveys, employing side-scanning laser beams to capture vast areas with a relative accuracy of 0.15 m and absolute accuracy of less than 0.5 m.
- **Ground-based LiDAR:** Used in urban or forested environments, it can be stationary or mobile and is applied in 3D modeling of infrastructure or natural ecosystems.
- **Single-Photon LiDAR (SPL):** Based on Time-Correlated Single-Photon Counting (TCSPC), this technology enables extreme sensitivity and high surface resolution, even in low photon return conditions (e.g., underwater imaging, long-range scanning).

Recent advancements in LiDAR systems have significantly increased pulse repetition rates to over 100 kHz, enabling point densities exceeding 10 points/m^2, which are crucial for applications such as forest remote sensing, 3D infrastructure modeling, and autonomous vehicle navigation. Moreover, LiDAR systems operating in the shortwave infrared (SWIR, 1.4–3 µm) spectrum offer advantages in terms of eye safety and reduced solar interference [4, 5, 6].

To observe how LiDAR is incorporated in cutting-edge computing systems for real-world applications, Figure 2.4 shows a typical pipeline for real-time 3D reconstruction. In the figure, a LiDAR sensor captures depth information from a scene at a distance of 320 meters, while an RGB camera simultaneously takes visual reference information. These input streams are taken to a GPU in real time and produce 3D reconstruction with dense geometric and visual data. This merging offers accurate and dense spatial mapping, which is essential for uses such as autonomous navigation, remote surveillance, and virtual environmental simulation.

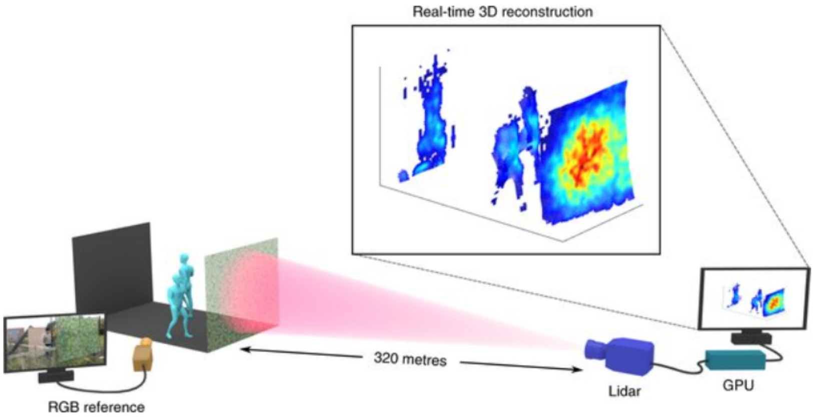


Figure 2.4: Real-time 3D reconstruction pipeline combining long-range LiDAR sensing and RGB reference imagery [7].

2.4.1 Evolution and Technological advancements of LiDAR systems

LiDAR (Light Detection and Ranging) technology has had an amazing progression since the conceptual foundations were laid in the early 20th century. The use of light to estimate distances began to take form in the 1930s, precisely in 1938 when light pulses were used to calculate cloud heights—marking one of the first practical applications of the principle. However, it was the 1960 development of the laser that actually kick-started the advancement of LiDAR as a coherent sensing technology.

The first commercial LiDAR system, Colidar (Coherent Light Detecting and Ranging), was made available in 1961 by Malcolm Stitch at Hughes Aircraft Company. Initially developed for military tracking applications, the Colidar Mark II (1963) was among the first land-based systems. NASA played a key role in expanding LiDAR's applications throughout the 1970s through the development of laser-based remote sensing techniques for environmental monitoring including ocean profiling and atmospheric measurements.

While the early LiDAR systems were very promising, their widespread application to high-accuracy applications was hindered until the mid-1980s by the absence of accurate positioning technologies. That is, until the development of Global Positioning System (GPS) and Inertial Measurement Unit (IMU) integration refined accurate geolocation, greatly stimulating imaging, mapping, and topographic analysis.

From the late 1980s, LiDAR demand in aerial photogrammetry and topographic surveys grew, stimulating further research and development. Commercial availability of precise positioning systems enabled LiDAR to be a powerful instrument in a wide range of applications from environmental mapping and forest canopy investigations to urban infrastructure planning and autonomous vehicle navigation [8, 9, 6].

Recent years have seen unparalleled expansion in LiDAR performance and flexibility, driven by developments in semiconductor and photodetector technology. Perhaps the most significant advance has been the mating of silicon-based Single-Photon Avalanche Diode (SPAD) detectors that are realized in complementary metal-oxide-semiconductor (CMOS) technology. Such SPADs achieve single-photon detection with extremely high sensitivity and at picosecond-scale time resolution [10].

Such temporal precision makes SPAD detectors an ideal fit for Time-Correlated Single-Photon Counting (TCSPC), which allows for ultra-fine spatial resolution even at extensive ranges. This makes SPAD-based LiDAR extremely well-suited for long-range and high-altitude use, for instance, airborne remote sensing [11].

Furthermore, modern LiDAR systems are increasingly coupled with powerful Graphics Processing Units (GPUs), enabling real-time data processing and 3D reconstruction. GPU acceleration enables the generation of very dense point clouds and 3D models in almost real-time, benefiting applications such as autonomous driving, robotics, and augmented reality.

From its beginnings in experimental techniques to its present position as the foundation of sophisticated sensing and mapping technologies, LiDAR has matured into a powerful and essential tool. Through continued research as well as engineering innovation, LiDAR remains at the forefront of redefining the limits of 3D perception, delivering unparalleled precision, speed, and environmental awareness to a wide range of scientific, industrial, and commercial applications [12, 13, 14].

2.4.2 LiDAR working principle

LiDAR (Light Detection and Ranging) operates based on the principle of measuring the time taken for a laser pulse to travel to a target and reflect back to the sensor. By analyzing this time-of-flight (ToF) data, LiDAR systems can calculate distances with high precision and generate detailed 3D maps of the surrounding environment.

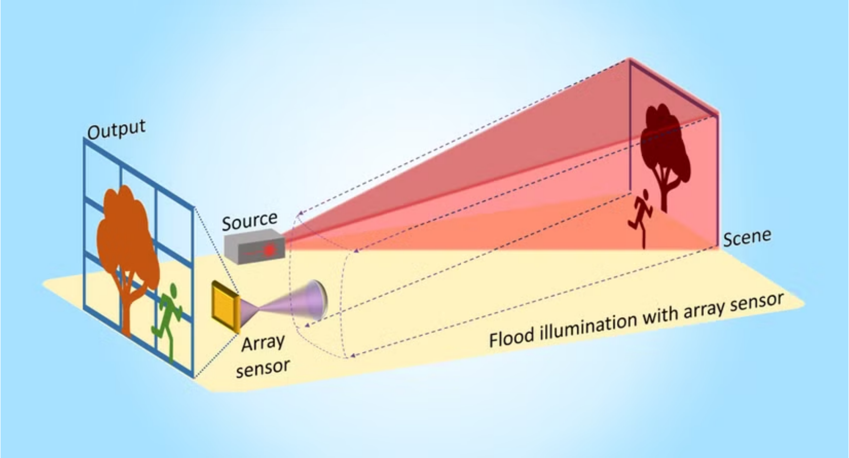


Figure 2.5: LiDAR working principle [15]

This figure 2.5 illustrates the operating principle of a 3D imaging system that utilizes flood illumination in combination with an array sensor—a configuration commonly employed in Time-of-Flight (ToF) or structured light technologies. The main components involved are:

- **Laser Source:** A short laser pulse (typically in the nanosecond range) is emitted toward the scene. The wavelength—often in the near-infrared range—is selected to ensure strong interaction with a variety of surfaces and resistance to atmospheric interference.
- **Light Propagation and Reflection:** The emitted beam spreads across the scene, illuminating all objects in its path. These objects reflect a portion of the light back toward the sensor.

- **Array Sensor:** The reflected light is captured by a 2D array of photodetectors. Each pixel in the array corresponds to a specific point in the scene, enabling fast and parallel acquisition of depth information across the entire field of view.
- **Output (3D Reconstruction):** The system calculates the time-of-flight (ToF) for each laser pulse—the time it takes to travel to the object and back. From this, it estimates the distance to each point, producing a 3D depth image or point cloud of the observed environment.

This type of LiDAR architecture is well-suited for compact, real-time systems such as those in smartphones, drones, or robotics, offering fast and wide-area 3D perception without moving parts.

A – Operational process of a LiDAR System:

As shown in the figure 2.6, power is supplied to the laser through the laser power supply. The laser beam then passes through a set of focusing optics before reaching the target object. The reflected laser light is collected through another set of focusing optics, where a specific wavelength is filtered by an optical filter. The filtered output is directed to a photodetector, which converts the optical signal into an electrical one. This data is then transmitted to a computer for analysis and visualization. The entire LiDAR system is mounted on a rotating mechanism, typically powered by a motor, allowing for continuous scanning and mapping of the environment .

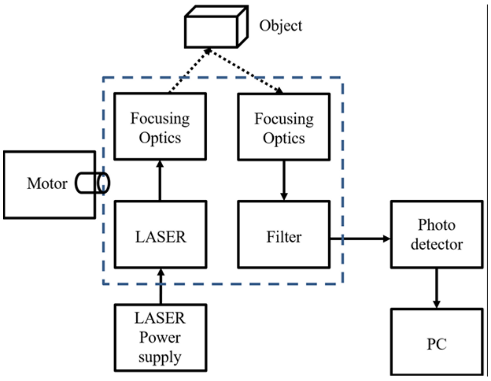


Figure 2.6: Operational block diagram of LiDAR [6]

B – Time-of-Flight (ToF) principle:

The fundamental principle behind LiDAR (Light Detection and Ranging) lies in Time-of-Flight (ToF) measurements, where laser pulses are emitted toward a target, and the time taken for the reflected light to return is used to calculate distance with high precision. As illustrated in Figure 2.7, the ToF process involves emitting short laser pulses, detecting their reflections with a photodetector, and analyzing the time delay to estimate distance. This method enables accurate 3D mapping of environments and is a cornerstone of modern LiDAR systems [16].
The system records the time interval (ΔT) between emission and reception, and the distance (D) to the object is determined using the formula:

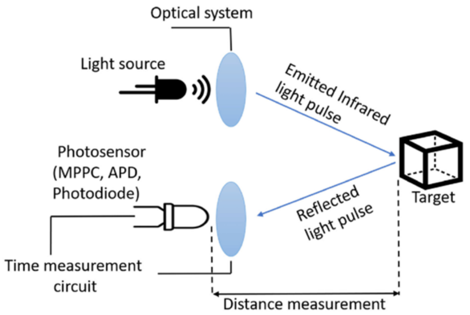


Figure 2.7: time of flight ToF principle [16]

$$D = \frac{c \cdot \Delta T}{2} \tag{2.1}$$

where:

- D = Distance to the target
- c = Speed of light ($\approx 3.0 \times 10^8$ m/s)
- ΔT = Time taken for the light to travel to the object and return

This method is widely used in pulsed LiDAR systems and allows for high-speed and accurate distance measurements.

C – Alternative measurement approaches:
While ToF pulsed lasers are the most common, two alternative techniques exist for measuring distances:

1. **Phase-Shift measurement (AMCW LiDAR)**

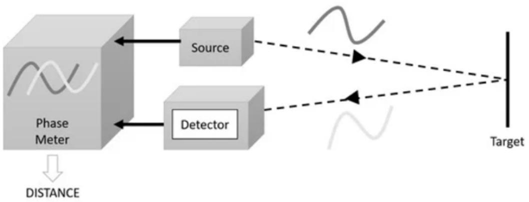


Figure 2.8: Time of flight phase-measurement principle used in AMCW sensors [17]

The figure 2.8 above illustrates the Phase-Shift Measurement method used in AMCW LiDAR operates based on the following principles :

- This approach uses an amplitude-modulated continuous wave (AMCW) laser.
- The phase difference ($\Delta\Phi$) between the emitted and received light is used to compute the distance.
- f_M denotes the modulation frequency.
- This method is highly effective for short-range applications, but it is limited by range ambiguity beyond approximately 100 meters.

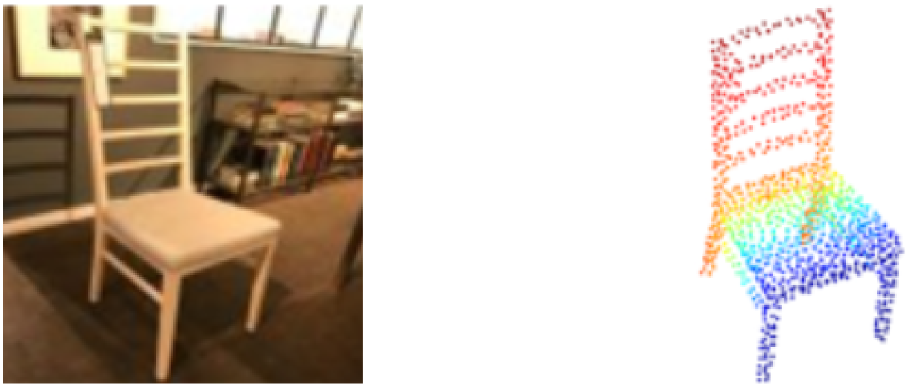
2. **Frequency modulated continuous wave (FMCW LiDAR)**
   Frequency Modulated Continuous Wave (FMCW) LiDAR systems operate based on the following principles:

   - The instantaneous frequency of the emitted laser is modulated over time.
   - The frequency shift between emitted and reflected light is analyzed to determine distance.
   - This method provides superior depth resolution (as low as 0.1 cm) and improved immunity to interference.

2.4.3 LiDAR Data quality

A – LiDAR data output and processing formats:

To better understand the structure of LiDAR data, it is helpful to begin with a visual example. Figure 2.9 compares a standard RGB image (a) with its corresponding 3D point cloud (b), where depth is color-coded: blue represents nearby surfaces, yellow indicates mid-range, and red denotes distant areas. This comparison illustrates how 2D visual input can be transformed into a structured 3D representation—an essential step in LiDAR-based perception and analysis.



(a) RGB image

(b) 3D point cloud

Figure 2.9: Example of an RGB image (a) and its corresponding point cloud (b), with color representing depth [18].

LiDAR systems generate point clouds, spatial datasets composed of 3D Cartesian coordinates (x, y, z) that precisely capture the geometry of real-world scenes. The initial data output—raw range measurements and sensor orientation—is transformed through calibration and coordinate conversion into a coherent 3D representation within a local or global reference frame.

Modern systems often enrich these point clouds with ancillary information such as return intensity, return number, timestamp, and RGB values from co-registered cameras. This multidimensional data enhances the spectral and spatial fidelity of the representation, enabling more accurate analysis.

However, raw point clouds are typically sparse and irregular, limiting their direct usability in tasks like object detection or 3D reconstruction. To overcome this, deep learning-based upsampling techniques such as PU-Net [19], PU-GCN [20], and RS-CNN have been developed to generate denser, more uniform point distributions while preserving fine geometric details.

As LiDAR hardware and computational capabilities have advanced, point clouds have evolved from static representations to dynamic, high-resolution datasets. Real-time processing, predictive modeling, and integration with AI are now feasible thanks to GPU acceleration and improved storage solutions. Consequently, point clouds form the backbone of digital twins, smart environments, and a wide array of emerging applications across science, engineering, and industry [21, 22, 23].

In the majority of state-of-the-art LiDAR systems, particularly dToF or TCSPC-based systems, the photon count temporal histogram is the fundamental measurement from which depth information is extracted. The histogram records the count of photon arrivals received over discrete time bins and provides a statistical profile of return times per detection unit or pixel. The shape and peak of the histogram directly encode the scene's depth structure.

Raw histogram data is post-processed after collection—often with matched filters, deconvolution, or deep learning networks—to infer depth, and then geometrically transformed using sensor orientation and calibration parameters to generate 3D point clouds in global or local coordinate frames.

This end-to-end pipeline—from photon-counting histograms to dense, semantic 3D models—enables important applications in autonomous navigation, remote sensing, digital twins, and smart infrastructure systems.

B – Photon count and signal quality in ToF imaging (PPP):

The Photons Per Pixel (PPP) value is the mean number of photons received per pixel within one acquisition period in a time-of-flight (ToF) camera system. PPP is an absolute measurement of the intensity of the arriving signal at the sensor and is a function of the reflectance of the scene, illumination power, exposure time, and the efficiency of the optical system. A higher PPP is linked with a higher number of detected photons, which reduces the quantum (Poisson) noise statistically and leads to more stable and accurate depth estimation. On the other hand, low PPP yields a sparse photon distribution across time bins, increasing uncertainty and lowering the signal-to-noise ratio (SNR). This is common in low-light environments or when viewing distant or low-reflectivity surfaces. PPP consequently plays an important part in defining the robustness and quality of the histogram generated by the ToF sensor and the performance of any subsequent depth reconstruction algorithm [12].

C – Signal-to-Background Ratio (SBR) in depth estimation:

The **Signal-to-Background Ratio (SBR)** calculates the ratio of signal photons—those that travel directly from the light source, are reflected from scene surfaces, and reach the sensor—to the background photons, which may originate from ambient light, multiple scattering, or indirect reflections.

It is defined as:

$$\text{SBR} = \frac{\text{Sum of signal counts}}{\text{Sum of Background counts}} \tag{2.2}$$

A large SBR indicates that most of the detected photons are signal photons, which results in well-defined, localized peaks in the histogram and consequently more accurate depth estimation.

However, a small SBR indicates severe contamination by background noise, which can spread the temporal profile of the histogram and compromise the reliability of depth reconstruction methods—particularly in the case of outdoor scenes or complex optical environments [12].

D – Complementary roles of PPP and SBR:

PPP and SBR both determine the quality of the signal and the precision of the depth sensing of a ToF system. High PPP assures statistical stability, while high SBR makes detected photons informative and useful. Low values for either of them can lead to noisy or uncertain histograms, making depth estimation highly sensitive to the reconstruction algorithm used. For instance, simple techniques such as argmax will fail in low SBR or low PPP conditions, whereas robust techniques such as matched filtering will still provide valuable output by accommodating anticipated signal profiles [12].

2.4.4 Temporal Histogram simulation for dToF sensors

During this work, the temporal histogram of a single pixel is generated to model how a real direct time-of-flight (dToF) sensor performs under mixed signal and noise conditions. Signal strength as well as background contamination are modeled by two governing parameters: **Photons Per Pixel (PPP)** and **Signal-to-Background Ratio (SBR)**.

The process begins by dividing the total number of photons per pixel (PPP) between the signal component (Lev_S) and the background component (Lev_B), according to the SBR as follows:

$$\text{Lev}_S = \frac{\text{PPP} \times \text{SBR}}{1 + \text{SBR}}, \quad \text{Lev}_B = \text{PPP} - \text{Lev}_S$$

Then, for each pixel, the signal shape is approximated as a Gaussian distribution centered at the pixel's actual depth value. Its amplitude is modulated by the pixel's reflectivity, computed from the RGB image using the luminance formula:

$$Y = 0.299R + 0.587G + 0.114B$$

The obtained grayscale reflectivity map is normalized to lie in the interval $[0, 1]$. The depth range is discretized into T temporal bins (e.g., $T = 300$), and the normalized depth value for each pixel is computed to determine the center of the Gaussian.

Each pixel is modeled by a histogram with a weighted Gaussian-shaped signal (with the true depth as center and scaled by reflectivity and Lev_S) superposed onto a uniformly distributed background noise term in every bin:

$$H(t) = r_0 \cdot \text{Lev}_S \cdot G(t, d_0) + \frac{\text{Lev}_B}{T}$$

where $G(t, d_0)$ is the normalized Gaussian function centered at the pixel's true depth d_0, and r_0 is the pixel's normalized reflectivity.

Finally, to simulate realistic sensor behavior, Poisson noise is added to the summed histogram values. This process simulates the statistical nature of photon arrival in actual sensing conditions and introduces uncertainty that is directly proportional to PPP.

Using this simulation strategy, we generated noisy depth maps under different sensing conditions by varying the values of PPP and SBR. Three configurations were experimented:

- **PPP = 1, SBR = 0.25:** noise-saturated and low-light environment,
- **PPP = 4, SBR = 1:** signal and background are equal,
- **PPP = 16, SBR = 4:** ideal, high-signal conditions.

Figure 2.10 presents examples of recorded photon histograms under various sensing conditions.

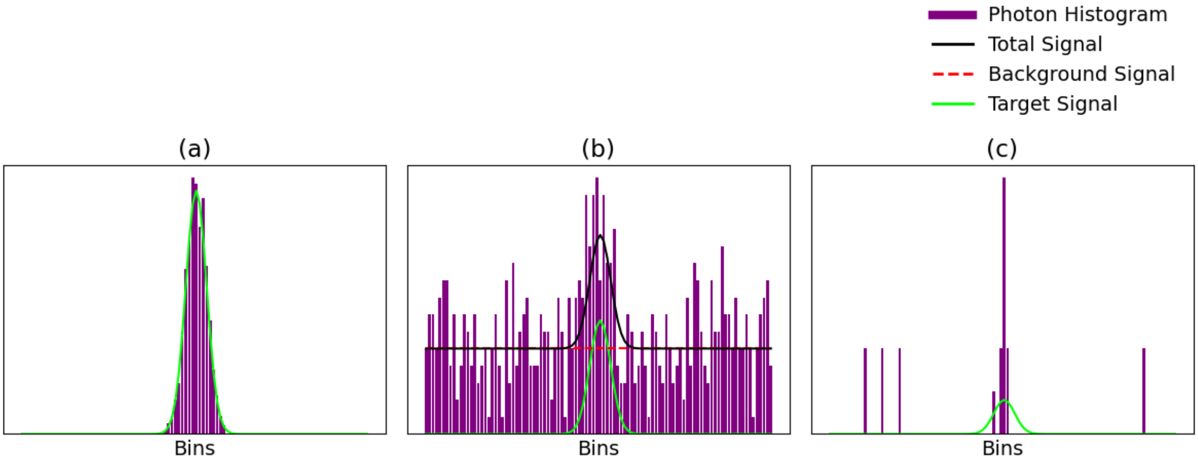


Figure 2.10: Photon histogram simulation under different sensing conditions : (a) the ideal case, (b) strong background illumination, (c) few photons.

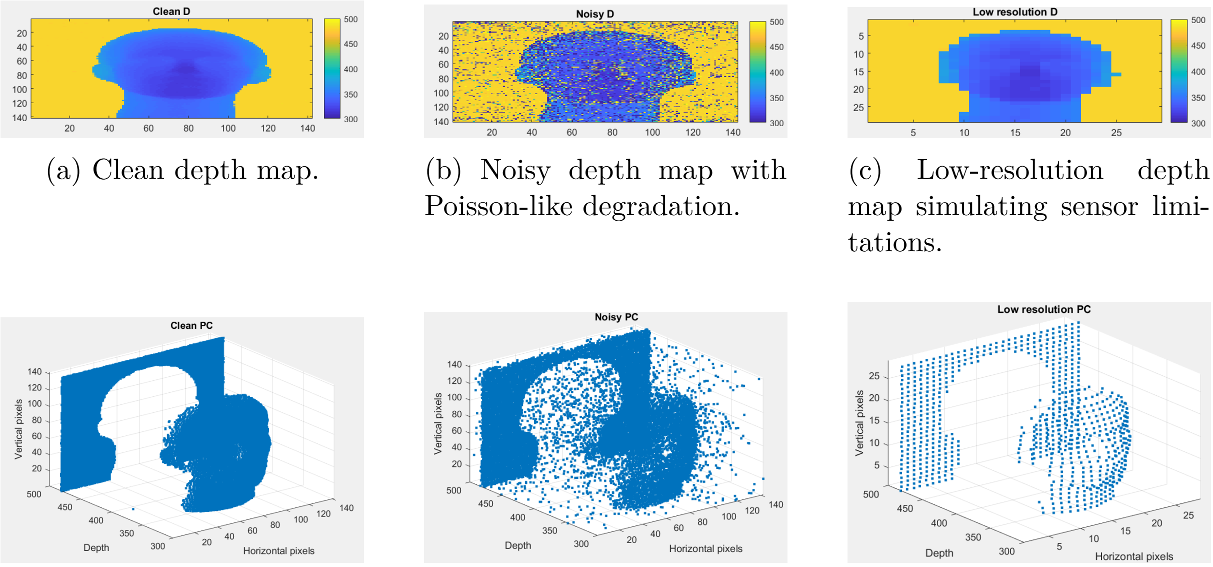
- **Subfigure (a)** illustrates the ideal case, where the photon signal is well-defined with minimal background noise, leading to a sharp Gaussian peak.
- **Subfigure (b)** shows the effect of strong background illumination (low SBR), where the histogram becomes irregular and noisy, blurring the peak.
- **Subfigure (c)**corresponds to a scenario with few photons (low PPP), resulting in a sparse and poorly defined histogram.

These cases demonstrate how signal clarity degrades with decreasing photon count and increasing background noise, directly impacting depth estimation accuracy.

2.4.5 Challenges and issues

Although 3D LiDAR offers impressive capabilities, it still faces a number of technical and practical challenges. One of the first and most obvious hurdles is the cost. High-quality LiDAR systems especially the ones capable of detailed, high resolution scanning aren't cheap. They require precise laser components, sensitive detectors, powerful processors, and sometimes even moving parts. That means both buying and maintaining them can be expensive, making them less practical for smaller scale or budget-conscious projects.

Another critical issue is data quality. While LiDAR is highly precise in theory, the data collected is often affected by various sources of noise and degradation. Environmental factors such as dust, fog, or highly reflective surfaces (like glass or water) can scatter the laser pulses or distort returns. Additionally, sensor limitations such as low spatial resolution—common in lightweight or embedded LiDARs—further impact the quality of the depth information.



(a) Clean depth map.

(b) Noisy depth map with Poisson-like degradation.

(c) Low-resolution depth map simulating sensor limitations.

(d) 3D point cloud from clean depth.

(e) Noisy 3D point cloud affected by degraded depth input.

(f) Sparse 3D point cloud from low-resolution depth.

Figure 2.11: Illustration of depth map degradations and their effects on 3D reconstruction. In the point clouds, **blue indicates closer surfaces** and **yellow more distant areas**.

To further illustrate the impact of data degradations on 3D perception, Figure 2.11 presents a comparative analysis of depth maps and their resulting point clouds under various conditions:

(a) shows an ideal, clean depth map;

(b) introduces Poisson-like noise, simulating realistic measurement degradation;

(c) represents a low-resolution depth map, mimicking hardware limitations;

(d) is the 3D point cloud derived from (a), offering high spatial fidelity;

(e) shows the effect of noise on 3D reconstruction, with increased distortion;

(f) displays a sparse and less informative point cloud due to reduced resolution.

These examples clearly demonstrate how input quality directly influences the geometry and density of 3D reconstructions. In the visualizations, blue indicates nearby surfaces, while yellow represents more distant regions, offering intuitive depth perception. The degradation scenarios emphasize the need for robust processing techniques capable of handling noisy or incomplete LiDAR data.

Another challenge lies in the range–resolution trade-off: increasing range often comes at the expense of spatial detail, which is critical in applications like autonomous driving or aerial mapping.

While LiDAR performs well in low-light conditions, adverse weather (rain, snow, fog) can scatter laser beams and reduce effectiveness. Certain materials—dark or reflective—also yield poor returns.

Furthermore, LiDAR generates large volumes of data, requiring substantial storage and computing resources, particularly for real-time applications.

Mechanically scanned LiDAR systems involve moving parts that wear out over time. Solid-state LiDAR, a more robust alternative, is promising but not yet widely adopted. High power consumption further limits deployment on mobile or battery-powered platforms.

Finally, LiDAR is often integrated with cameras, GPS, and IMUs. However, multisensor fusion is complex due to differences in resolution, frame rates, and fields of view. Synchronization and calibration demand advanced algorithms, and the lack of standardization hinders interoperability. Moreover, as LiDAR becomes more widespread in public spaces, it raises privacy concerns, particularly in surveillance and behavioral monitoring. [9].

2.5 Conclusion

This first chapter focused on the theoretical foundation and key issues inherent with computational imaging, with special consideration to the application of LiDAR sensors in this context. LiDAR, through its ability for precise and dense depth sensing, has become a key component of modern 3D imaging systems. However, its weaknesses most notably in the areas of spatial resolution, noise, and environmental sensitivity have served to increase interest in computational imaging techniques.

Computational imaging attempts to overcome the limitations of physical sensors using physical models and advanced algorithms, generally based on deep learning. The addition of LiDAR data enables the combination of reliable physical measurements and powerful reconstruction algorithms, paving the way for robust, precise, and efficient 3D imaging solutions.

CHAPTER 3 :

State-of-the-art

In this chapter, we provide an overview of existing methods for multimodal 3D imaging, covering both traditional techniques and data-driven approaches. The chapter is structured in two parts: the first focuses on the core architectures and representative models developed for key tasks such as super-resolution and multimodal 3D imaging; the second part is dedicated to UC-Net [24], a segmentation model that plays a central role in our experimental framework.

3.1 Existing methods for multimodal 3D imaging :

Multimodal sensors capture and integrate diverse characteristics of a scene to maximize information gain. In optics, this may involve capturing intensity in specific spectra or polarization states to determine factors such as material properties or an individual's health conditions. Combining multimodal camera data with shape data from 3D sensors is a challenging issue. Multimodal cameras, e.g., hyperspectral cameras, or cameras outside the visible light spectrum, e.g., thermal cameras, lack strongly in terms of resolution and image quality compared with state-of-the-art photo cameras. These are some existing methods for multimodal :

3.1.1 Traditional approaches for multimodal image fusion :

Traditional approaches for multimodal image fusion refer to a set of techniques and methods developed over the years for integrating information from multiple sources of images. These approaches are typically based on mathematical or statistical models and involve extracting features from the input images and fusing them to generate a single output image.

As shown in Figure 3.1,**Transform-based fusion, dictionary-based fusion, and statistical based fusion** are examples of traditional approaches that have been widely used in the literature. These traditional approaches have been used in various medical imaging, remote sensing, surveillance, and industrial image processing applications.

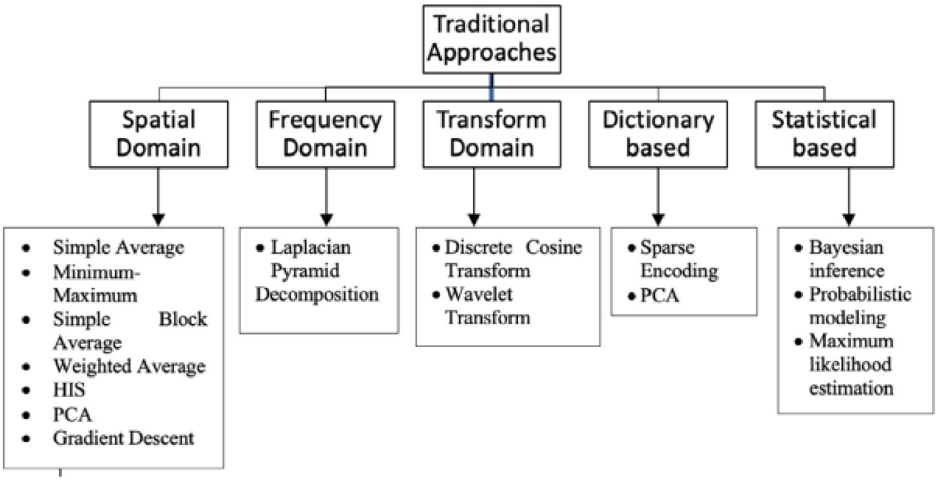


Figure 3.1: Classification of traditional multimodal image fusion approaches. [25]

In the following, we will discuss some of these traditional approches :

A – Transform-Based Fusion: Transform domain techniques involve converting images from the spatial domain into a different representation—typically the frequency domain—before fusion. This approach facilitates more effective integration of image details across various scales and resolutions. Among the commonly used methods in this category are:

- **Discrete Cosine Transform (DCT):** A widely used method, particularly in image compression (e.g., JPEG). DCT helps isolate image features in the frequency domain, and in the context of image fusion, DCT coefficients from different modalities are selectively combined to enhance relevant information while suppressing redundancy.
- **Wavelet Transform:** Another popular method that decomposes images into sub-bands representing different frequency components. It supports multi-resolution analysis, allowing detailed information from each source image to be fused at the appropriate scale. This results in high-quality fused images with well-preserved texture and contrast.

B – Dictionary-Based Fusion:
Dictionary-based methods use a learned or predefined set of basis functions (called a "dictionary") to represent image patches in a sparse or compact form. These techniques are powerful for preserving structural details and removing noise. Among the commonly used methods in this category are:

- **Sparse Encoding:** Each image patch is represented as a sparse linear combination of dictionary atoms. Fusion is performed by selecting or averaging the sparse representations to form a more informative fused representation. This method excels in retaining fine details and edges.

- **Principal Component Analysis (PCA):** Although PCA is primarily a statistical tool, it is often used within dictionary-based approaches. It transforms correlated image data into a set of orthogonal components (principal components), and the fusion is typically carried out on the first few components that capture the most significant variance (i.e., information).

C – Statistical-Based Fusion: : Statistical-based image fusion techniques rely on mathematical models and probability theory to combine information from multiple modalities. These methods aim to model the uncertainty and variability inherent in the imaging process, often leading to more robust and reliable fusion outcomes. Key approaches include Bayesian inference, where prior knowledge and observed data are integrated to estimate the most probable fused image; probabilistic modeling, which uses statistical tools such as Gaussian Mixture Models to represent data distributions and manage noise; and maximum likelihood estimation (MLE), which identifies the fusion parameters that maximize the probability of observing the input images. These techniques are particularly effective in medical image fusion and remote sensing applications where uncertainty is significant.

3.1.2 Data-Driven Methods

The motivation for introducing deep learning into image fusion is to overcome the limitations of traditional methods. Deep learning has become a popular technique for multimodal image fusion due to its ability to automatically learn complex mappings between different modalities and efficiently handle large amounts of data.

In the following, we first present some key architectures used in data-driven methods, and then examine several existing models.

Common Deep Learning Architectures :

1 - Convolutional Neural Networks CNNs :

Convolutional Neural Networks (CNNs) become a central technology for image processing and increasingly applied to multimodal image fusion. Compared with traditional approaches, CNNs can automatically learn hierarchical spatial features from input images, which are particularly well-behaved to tap complementary and salient information from different modalities.

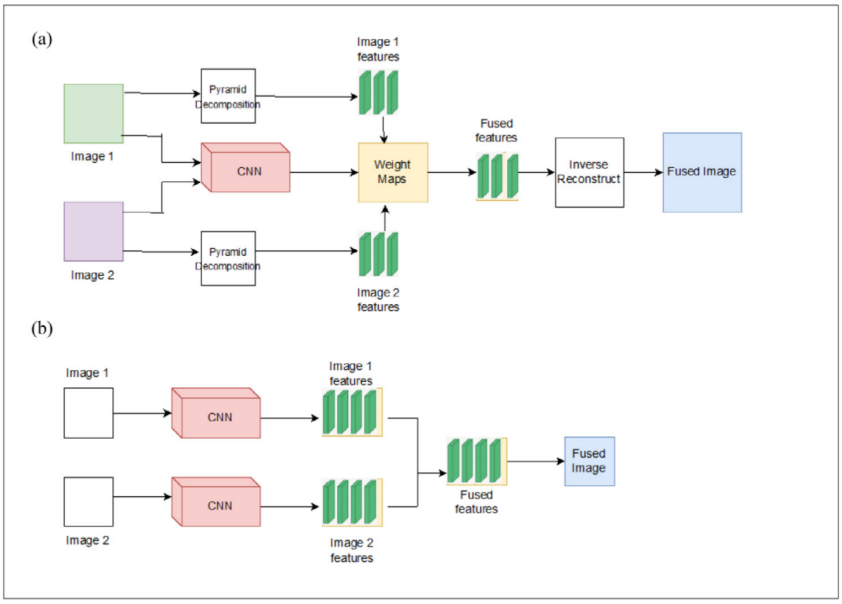


Figure 3.2: CNNs architucture for image fusion [25]

Typical steps involved in using CNNs for multimodal image fusion are shown in the figure 3.2 :

1. **Data Preparation** The input images are pre-processed to ensure consistency. This involves resizing them to a common size, normalizing pixel values, and organizing image channels appropriately.

2. **CNN Architecture Selection** Select an appropriate convolutional neural network based on the fusion goal. Popular choices include VGG for its simplicity, ResNet for deep feature extraction, and Inception for handling multi-scale features.

3. **CNN Training** The selected CNN is trained on datasets of paired images and their corresponding fused images. A loss function guides the training to preserve key structural and semantic details from both input images.

4. **Image Fusion Process** Once trained, the CNN takes new image pairs, extracts features from each, and combines them to form a fused representation that integrates the most relevant information from both.

5. **Evaluation and Refinement** The fused results are evaluated using quantitative metrics like PSNR and SSIM, along with visual assessment. Based on performance, the model may be fine-tuned for better quality fusion.

2 - Multimodal image fusion using Auto-encoders :

Auto-encoders are another type of deep learning model that can be used for multimodal image fusion. Auto-encoders can be used to perform feature-level fusion, where the input images are encoded into a lower-dimensional feature space and then decoded to create a fused representation; the figure 3.3 shows the general architecture of Multimodal Image Fusion using Stacked Auto-encoders.Here are the general steps :

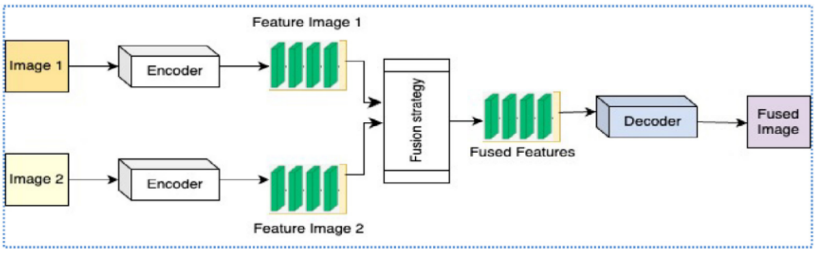


Figure 3.3: Auto-encoder architecture for image fusion [25]

1. **Data Preparation**
   Input images are pre-processed—resized to a common size, normalized, and channels are separated if needed.

2. **Auto-Encoder Architecture Selection**
   An appropriate auto-encoder is chosen, typically with a convolutional encoder and a deconvolutional decoder for image reconstruction.

3. **Training the Auto-Encoder**
   The model is trained on image pairs and their fused targets using a loss function that combines MSE and SSIM to retain important information from both inputs.

4. **Image Fusion**
   After training, the encoder extracts features from each image, which are then combined and passed through the decoder to generate the final fused image.

3 - Multimodal Image Fusion Using Generative Adversarial Networks :
Generative Adversarial Networks (GANs) [26] are a class of deep learning models consisting of two neural networks — a generator and a discriminator — trained in opposition. The generator aims to produce realistic data, while the discriminator attempts to distinguish between real and generated data, thus encouraging the generator to create increasingly convincing outputs. Initially developed for image synthesis, GANs have also been applied to multimodal image fusion, where the goal is to generate a single fused image by combining salient features from multiple input modalities. This process is illustrated in Figure 3.4.

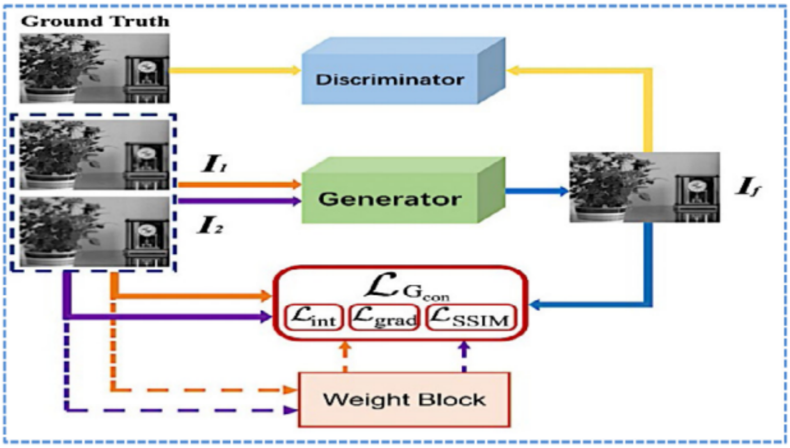


Figure 3.4: GAN architecture for Multimodal image fusion [25]

1. **Data Preparation**
   Input images are resized, normalized, and formatted (e.g., channel separation) to be compatible with the GAN model.

2. **GAN Architecture Selection**
   A suitable GAN architecture is chosen, typically with a convolutional-deconvolutional generator and a discriminator to differentiate real and generated fused images.

3. **Training the GAN**
   The generator learns to produce fused images similar to ground truth, while the discriminator learns to tell real from fake. A combination of adversarial and content loss (e.g., SSIM, MSE) ensures realistic and informative fusion.

4. **Image Fusion**
   After training, the generator fuses new image pairs by extracting and merging their features to produce a single fused image.

4 - Multimodal image fusion using transformers :
Transformers are a type of deep learning model widely used in natural language processing (NLP), but they can also be applied to image fusion tasks. Attention mechanisms can be used in multimodal image fusion to enable the model to focus on the most relevant features from each input modality.

Different attention mechanisms can be used for fusion tasks. Currently, Transformer architectures are not much explored for multimodal image fusion. Here is an overview of how Transformers can be applied to multimodal image fusion , the figure 3.5 shows the general architecture of the model :

1. **Understanding Multimodal Image Fusion**
   Multimodal image fusion involves combining information from multiple sources or modalities (e.g., visible light, infrared, depth) to create a single, more informative image. Each modality typically provides unique information, and the goal is to fuse this information to enhance the overall image quality or extract specific features.
2. **Transformers in Image Fusion**
   Transformers are robust neural network architectures that have shown remarkable success in various tasks due to their ability to capture complex dependencies in data. In image fusion, Transformers can be used to learn the relationships and dependencies between the different modalities and create a fused representation.
3. **Input Representation**
   The input to a Transformer-based multimodal image fusion system includes multiple source images from different modalities. Each source image is usually passed through a convolutional neural network (CNN) to extract image features. These features are then combined and passed as input to the Transformer model.
4. **Attention Mechanism**
   Transformers leverage attention mechanisms to weigh the importance of different parts of the input features when making predictions. In the context of image fusion, the attention mechanism can highlight regions or features from each modality that are most relevant for fusion.

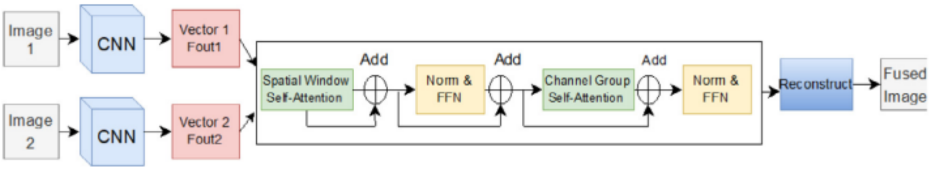


Figure 3.5: Image fusion using Transformers [25]

Overview of existing data-driven models

One of the most notable contributions in this area is the work titled *Consistent Direct Time-of-Flight Video Depth Super-Resolution* [27], which addresses the super-resolution problem by fusing low-resolution dToF measurements with high-resolution RGB images. Unlike traditional per-frame RGB-guided depth enhancement methods, the authors propose the first multi-frame fusion approach, introducing two novel models:

- **Depth Video Super-Resolution (DVSR)**: Exploits multi-frame correlations to enhance geometry prediction and temporal consistency, achieving superior performance compared to state-of-the-art per-frame methods [28], while maintaining a lightweight architecture.
- **Histogram Video Super-Resolution (HVSR)**: Further incorporates unique dToF histogram information, which reduces spatial ambiguity and flying pixels, thereby improving geometric fidelity.

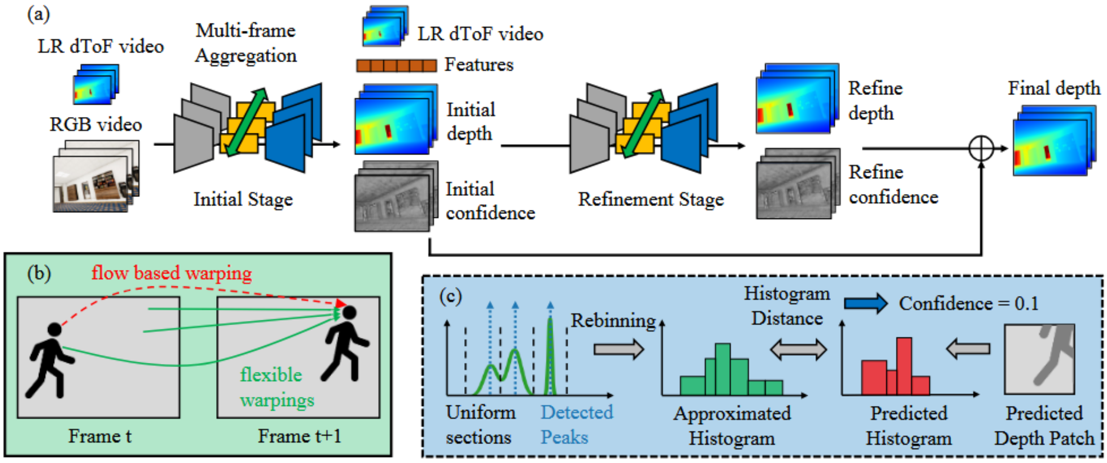


Figure 3.6: Proposed dToF video super-resolution framework. [27]

Figure 3.6 shows the proposed dToF video super-resolution framework. It generally follows a two-stage prediction strategy, where both stages predict a depth map and a confidence map that are fused to obtain the final prediction. Features are aligned and aggregated between frames, either bidirectionally or forward-only. (b) Schematic of flexible warping-based multi-frame feature aggregation. Instead of strictly following the estimated optical flow, features from multiple candidate positions are warped between frames. (c) Schematic of proposed histogram processing pipeline. The full histogram is compressed with peak detection and rebinning to produce an approximated histogram. At the confidence prediction stage, histogram distance is computed between the input histogram and the histogram generated by predicted depth values to estimate confidence in the prediction.

In parallel, significant work has been carried out on point cloud upsampling, which is closely related to depth map refinement. Notable examples include:

- **PU-Net** [19]: An upsampling framework that uses a hierarchical feature learning mechanism to progressively refine point representations. It employs interpolation-based restoration and introduces both reconstruction and repulsion losses during end-to-end training to produce more uniform and accurate point distributions. PU-Net addresses the challenge of upsampling irregular, unordered 3D point clouds, where traditional interpolation fails to ensure both geometric fidelity and distribution uniformity. The authors demonstrate that PU-Net outperforms prior optimization-based and learning-based approaches in terms of surface accuracy and uniformity, evaluated through metrics like Earth Mover's Distance (EMD) and Normalized Uniformity Coefficient (NUC).

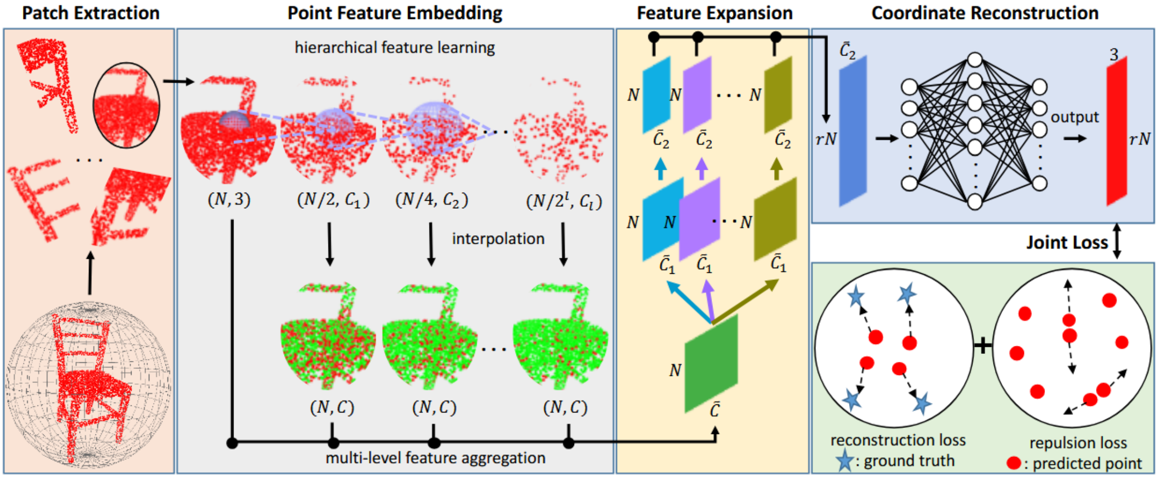


Figure 3.7: The architecture of PU-Net (better viewed in color) [19]

Figure 3.7 shows the architecture of PU-Net where input has N points, while the output has rN points, where r is the upsampling rate. C_i, $\tilde{C}$, and $\tilde{C}_i$ represent feature channel dimensions. Multi-level features are restored for the original N points using interpolation and then reduced to a fixed dimension C via a convolution. The red color in the point feature embedding component denotes the original and progressively subsampled points in the hierarchical feature learning process, while the green color indicates the restored features. A joint loss function combining reconstruction loss and repulsion loss is used in the end-to-end training of PU-Net.

- **PU-GCN** [20]: A point cloud upsampling method based on Graph Convolutional Networks (GCNs). It integrates the Inception DenseGCN multi-scale feature extractor with a novel NodeShuffle upsampling module. When incorporated into the 3PU [29] framework, PU-GCN improves structure preservation and fine detail reconstruction, effectively restoring features like the neck and ball shape of a faucet. PU-GCN addresses the limitations of prior MLP- or duplication-based upsampling techniques by leveraging graph convolutions to better encode local neighborhood structures and generate new points from learned features. Extensive experiments show that PU-GCN achieves state-of-the-art performance on benchmark datasets (e.g., PU-GAN's dataset and the newly introduced PU1K) with fewer parameters and faster inference, while preserving fine-grained details and reducing outliers.

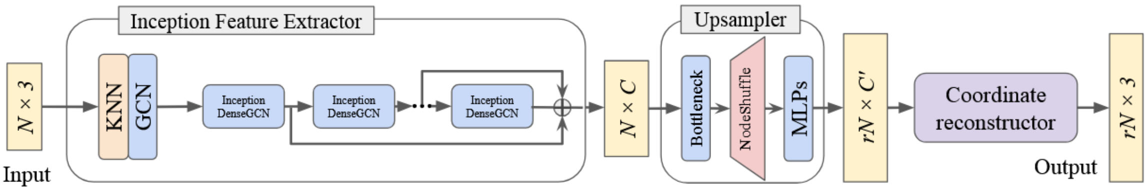


Figure 3.8: PU-GCN architecture. PU-GCN uses an inception feature extractor consisting of one or more Inception DenseGCN blocks, followed by the NodeShuffle based upsampler, and a coordinate reconstructor [20]

3.1.3 Hybrid Methods: Model-Based Deep Learning

By combining model-based and data-driven approaches, hybrid methods aim to leverage the advantages of both and mitigate their respective drawbacks. Rather than relying solely on purely data-driven techniques such as convolutional neural networks (CNNs) or generative adversarial networks (GANs), these approaches integrate prior knowledge, physical models, or mathematical optimization frameworks into deep learning architectures.

This integration leads to more interpretable, efficient, and robust models, particularly in tasks such as image fusion.

Advantages of Model-Based Deep Learning

- **Improved Interpretability:** These methods are grounded in established mathematical models, making their behavior easier to understand compared to purely data-driven networks.
- **Robustness in Challenging Conditions:** They tend to perform better under conditions such as noise, blur, or low lighting due to the incorporation of domain-specific knowledge.

To synthesize the various multimodal fusion strategies explored in this chapter—including traditional, deep learning, and hybrid approaches—we provide in Table 3.2 a comparative summary of representative methods. This table outlines the input modalities, architectural choices, fusion objectives, and evaluation metrics, offering a concise overview of current trends and positioning of our proposed framework.

Approach Type	**Principle**	**Advantages**	**Limitations**	**Examples**
Traditional	Based on physical/statistical models and rule-based fusion	Interpretable, requires little data	Limited adaptability, poor generalization	Kalman Filter, Bayesian fusion
Deep Learning	Data-driven models trained end-to-end for fusion tasks	High performance, automatic feature extraction	Requires large datasets, less interpretable	CNN-based fusion, GANs, Transformers
Hybrid (Model-Based DL)	Combines physical priors with learnable models	Better generalization, robust to noise, interpretable	Higher complexity, needs careful design	

Table 3.1: Comparison of multimodal fusion approaches in 3D imaging.

Table 3.2 provides a comparative overview of representative multimodal 3D image fusion methods presented in this chapter. It outlines the input modalities, network architectures, fusion mechanisms, and evaluation metrics, offering a clear perspective on current approaches and their respective strengths.

Method	**Modality Input**	**Architecture**	**Objective**	**Metric(s)**	**Year**
PU-Net [19]	Sparse Point Cloud	Hierarchical Feature CNN	Point cloud upsampling	Standard deviation, NUC	2018
GAN-based Fusion [26]	RGB, Depth	GAN (adversarial + content loss)	Image fusion	SSIM, MSE	2021
PU-GCN [20]	Sparse Point Cloud	Graph ConvNet (DenseGCN + NodeShuffle)	Point cloud upsampling	CD, HD, P2F	2021
Transformer Fusion [30]	RGB, Infrared, Depth	CNN + Transformer (attention)	Feature-level fusion	Attention scores	2023
DVSR [27]	Low-res depth maps + High-res RGB	Multi-frame CNN with alignment	Depth super-resolution	AE, TEPE	2023
HVSR [27]	ToF Histogram + High-res RGB	Histogram-aware CNN	Depth super-resolution	AE, TEPE	2023
Hybrid Fusion (Ours)	Tof Histogram + RGB	Multiscale + DVSR/HVSR + UC-Net	Depth super-resulotion + saliency detection	RMSE, AE	2024

Table 3.2: Comparison of state-of-the-art methods for multimodal 3D image fusion

To highlight the advantages of the proposed hybrid fusion method, Table 3.3 presents a comparison with the previous existing approaches based on key practical criteria.

Method	**Handles Noise**	**Low Photon Support**	**Multimodal Fusion**	**Saliency Detection**	**Realistic LiDAR Conditions**
PU-Net [19]	Limited	No	No	No	No
GAN-based Fusion [26]	Limited	No	RGB + Depth	No	No
PU-GCN [20]	Limited	No	No	No	No
Transformer Fusion [30]	Limited	No	RGB + Infrared + Depth	No	No
DVSR [27]	Limited	No	RGB + Depth	No	No
HVSR [27]	Limited	No	Histogram + RGB	No	Partial
Hybrid Fusion (Ours)	Yes	Yes	Histogram + RGB	Yes	Yes

Table 3.3: Comparison of existing methods with our proposed Hybrid Fusion approach.

3.2 High-level computer vision :

3.2.1 Saliency detection :

Saliency detection aims to identify the most visually important or attention-grabbing regions in an image or video—areas that are likely to attract human attention due to factors like contrast, color, shape, motion, or semantic content.

Existing RGB-D saliency detection methods treat the saliency detection task as a point estimation problem, and produce a single saliency map following a deterministic learning pipeline [24].However, this approach overlooks the inherently subjective nature of saliency perception, which can vary significantly between human annotators. To address this limitation, UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders [24] reformulates saliency detection as a distribution estimation task. Rather than generating a single output, UC-Net explicitly models the uncertainty in human annotations using a generative framework based on conditional variational autoencoders (CVAEs). This allows the network to capture a distribution over plausible saliency maps, better reflecting the variability and ambiguity inherent in human visual attention.

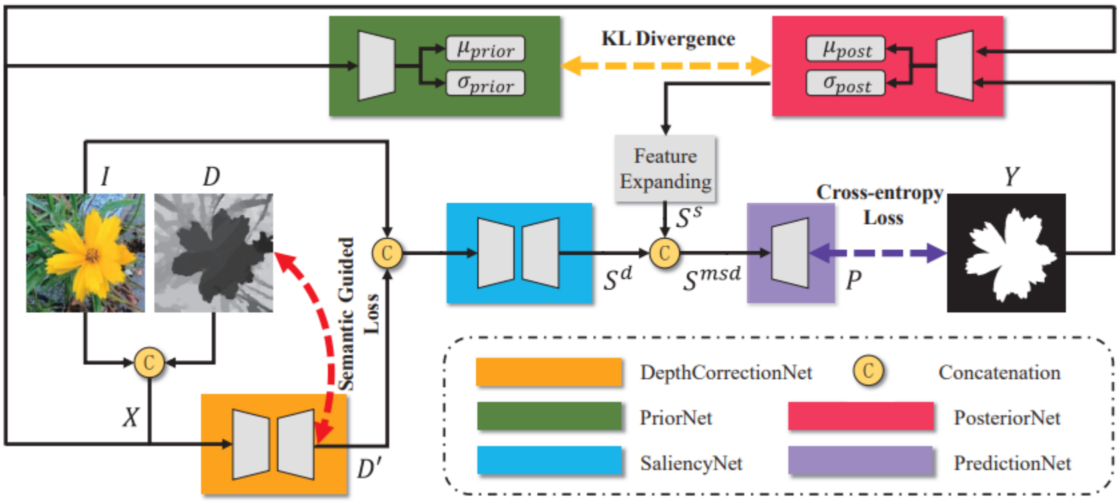


Figure 3.9: UC-NET. [24]

The figure 3.9 presents the UC-NET architecture, developed for RGB-D salient object detection (SOD). It utilizes both RGB images and depth maps to generate precise saliency predictions. The architecture will be detailed in the following section.

3.2.2 Architecture

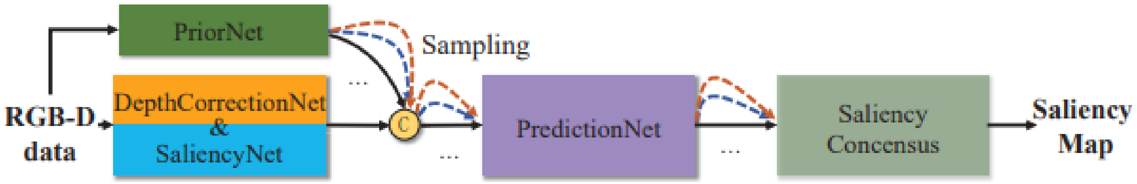


Figure 3.10: UC-Net framework [24]

The network is composed of five main modules, as illustrated in the figure 3.10:

1. **LatentNet (PriorNet & PosteriorNet)**
   It maps the RGB-D input X_i (for PriorNet) or X_i and Y_i (for PosteriorNet) to the low-dimensional latent variables $z_i \in \mathbb{R}^K$ (where K is the dimension of the latent space).

2. **DepthCorrectionNet**
   Takes the RGB image I_i and the depth image D_i, and refines the depth map into D_i', enhancing its quality and spatial consistency.

3. **SaliencyNet**
   Uses the RGB image I_i and the refined depth D_i' to extract deterministic saliency features S_i^d, which represent the core saliency cues.

4. **PredictionNet**
   Combines the deterministic features S_i^d from SaliencyNet and stochastic features S_i^s sampled from the latent space (via LatentNet) to predict a saliency map P_i.
   This sampling mechanism allows the network to generate diverse saliency predictions reflecting human uncertainty.

5. **Saliency Consensus Module (Testing Only)**
   Aggregates multiple stochastic predictions to produce a final consensus saliency map, mimicking the ground truth generation process used in datasets (which often aggregates multiple human annotations).

Building on the insights discussed previously, our research adopts a *Plug-and-Play* strategy for robust depth map fusion from RGB and LiDAR data, followed by the integration of a high-level computer vision task—saliency detection. Unlike conventional approaches, our focus lies in real-world environments affected by challenging degradation factors such as fog, smoke, sensor noise, and spatial misalignment.

While the original DVSR (Depth-Video Super-Resolution) and HVSR (Histogram-Video Super-Resolution) frameworks were mainly tailored for synthetic direct Time-of-Flight (dToF) datasets, we extend these concepts to more complex and realistic scenarios.

Our proposed method combines a statistical multiscale algorithm with a deep learning model -Depth Video Super-Resolution-, and concludes with saliency detection using the UCNet model.

By doing so, we aim to achieve high-fidelity, temporally consistent depth estimation suitable for practical applications in autonomous systems and AR/VR systems operating under adverse conditions. This work highlights the flexibility and generalizability of hybrid image fusion frameworks and underscores the practical potential of Plug-and-Play strategies in real-world deployments.

3.3 Conclusion

In this chapter, we reviewed traditional, deep learning, and hybrid methods for multimodal 3D imaging. While traditional approaches are interpretable and deep learning offers high performance, hybrid methods combine the strengths of both, making them a promising direction. These insights motivate the proposed method in the next chapter.

CHAPTER 4 :

Modeling and Proposed Approach

4.1 Introduction

As shown in the previous chapter “the state of the art”, a number of reconstruction techniques have been proposed to improve the quality of depth videos produced by Direct Time-of-Flight (dToF) sensors, which are accurate but limited by the low resolution. A novel solution has been suggested to overcome this limitation by taking advantage of two complementary neural network architectures:

- **DVSR (Depth Video Super-Resolution)** that generates temporal sequences of RGB and depth images to create more temporally coherent and sharper depth maps.
- **HVSR (Histogram-based Video Super-Resolution)**, an extension of this that also uses depth histograms from the sensor, enabling reconstruction of finer features more accurately.

For both training and testing their models, the authors supplied synthetic datasets (Replica, Dy-DToF) and a real-world dataset, ARKIT. The results demonstrate that their approach surpasses state-of-the-art methods in reconstruction quality and temporal stability. However, real-world sensor measurements generally have severe defects. Specifically, dToF sensors are susceptible to various types of noise, such as thermal noise, multipath reflections, and ambient light interference. Moreover, environmental aspects like illumination changes, transparent or reflective objects, and dynamic conditions (e.g., motion in the scene or from the sensor) can greatly affect the depth measurement accuracy.

These constraints render real-world data far less "clean" than synthetic data or data recorded in controlled environments. Therefore, a model trained on nothing but flawless data might not generalize when transferred to real-world environments. It is thus required to generalize and strengthen the algorithm by subjecting it to noisy data during training. This kind of generalization enables more realistic simulation of sensor behavior in real-world conditions. It aims at enhancing the robustness of the model to perturbations while achieving high reconstruction accuracy even in complicated and uncontrolled conditions.

4.2 Proposed approach and model architecture

As already discussed, our prime goal is to extend the HVSR and DVSR algorithms to noisy data, in order to more accurately replicate the real environment in which dToF sensors are deployed. To realize this goal, we have followed a systematic approach, as illustrated by the figure 4.1 below. There are five general steps in this method:

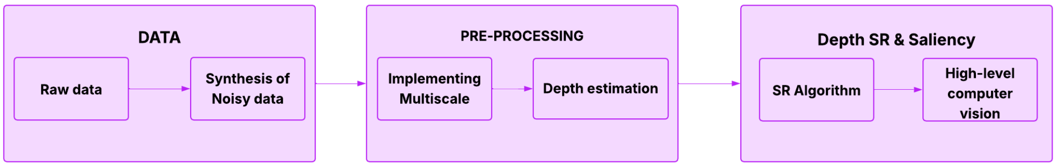


Figure 4.1: Data processing pipeline

DATA

1. **Synthesis of Noisy Data:** From raw input data, we synthesize depth data by introducing realistic noise and distortions that emulate the characteristics of real-world depth sensors.

PRE-PROCESSING

2. **Multiscale Approach:** Utilizing multiscale processing to obtain data at different spatial and temporal resolutions, improving the model's capability for extracting useful features even under noisy conditions.

3. **Depth Estimation:** The depth estimation process relies on Matched Filtering and ARGMAX techniques, which help suppress noise while preserving fine structural details, resulting in a more accurate reconstruction of the depth map.

Depth super resolution & Saliency

4. **Super-Resolution (DVSR & HVSR):** Running the algorithms on the simulated data to see how they perform under unfavorable conditions. This phase quantifies their robustness and their performance in generating coherent high-quality depth maps under realistic conditions. By doing so, we aim to validate the applicability of DVSR and HVSR through a procedure of confronting the models with data reproducing realistic conditions.

5. **High-level computer vision:** Seeks to highlight visually important regions in an image. In our work, we applied the UC-Net model to the data we generated, enabling uncertainty-aware saliency detection adapted to real-world conditions.

4.2.1 Noisy Data Generation

In this project, a 3D histogram for each image was constructed based on a merged RGB image and its corresponding depth map. The goal was to replicate what a time-of-flight (ToF) sensor would achieve by simulating significant physical phenomena such as:

- surface reflectance,
- actual depth values,
- signal-to-background ratio (SBR),
- average number of photons per pixel (PPP).

Reflectivity from the RGB image was approximated using the luminance formula:

$$Y = 0.299R + 0.587G + 0.114B \tag{4.1}$$

which converts the color image to a grayscale intensity representation. The resulting reflectivity map was normalized to the interval $[0, 1]$.

Depth values from the map were discretized into T uniform intervals (e.g., $T = 300$), where each interval corresponds to a temporal bin in the simulated ToF sensor response. The result was a 3D histogram, denoted as `z_values`, representing the photon count distribution across depth bins per pixel.

Once the histogram was generated, the depth map could be retrieved through two mechanisms:

- **Argmax:** selecting the depth bin with the highest photon count.
- **Matched filter approach:** correlating the histogram with a Gaussian kernel to estimate depth.

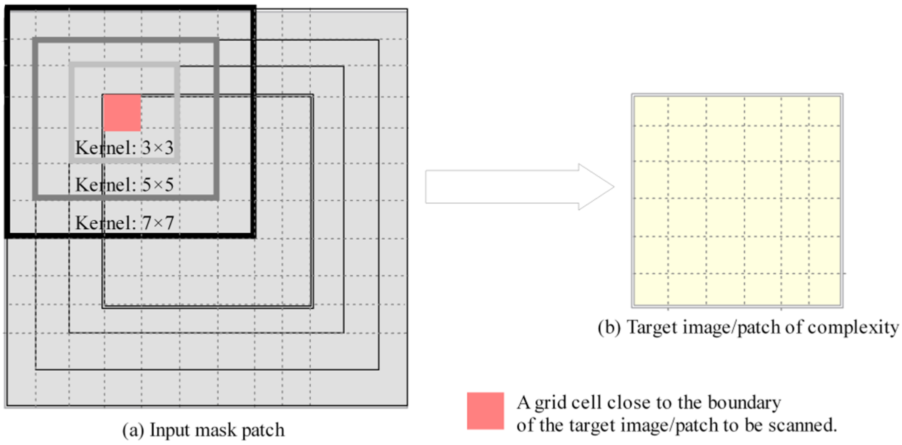
This pipeline allows us to simulate a virtual dToF sensing process and to evaluate the accuracy of reconstruction using objective metrics such as the Root Mean Square Error (RMSE).

4.2.2 Implementing multiscale [31]

A – Definition of Multiscale:

Multiscale processing is a technique that filters or analyzes a signal at different spatial scales to extract both fine-grained details and large-scale structural information. In this work, the technique is applied to the photon time histograms of each pixel using mean filters with increasing kernel sizes (e.g., 3×3, 5×5, etc.). This allows the attenuation of noise due to stochastic photon limitations while preserving local depth variations.

Figure 4.2 illustrates the multiscale convolution operator applied with different kernel sizes. This hierarchical processing enables the integration of information across multiple scales, leading to more robust and accurate depth reconstructions by balancing noise suppression with detail preservation.


Kernel: 3×3
Kernel: 5×5
Kernel: 7×7
🟥 A grid cell close to the boundary of the target image/patch to be scanned.

(a) Input mask patch

(b) Target image/patch of complexity

Figure 4.2: Multiscale convolution operator using different kernel sizes [32].

Multiscale processing integrates data across these different scales to create stronger and more robust depth reconstructions, with an optimal balance between noise reduction and preservation of detail.

B – Applying Multiscale Processing:

For reconstructing depth from noisy histogram-type measurements, the presence of noise can significantly affect the accuracy of the recovered depth, especially when the number of photons per pixel (PPP) is low.

To counteract such effects, an efficient strategy consists in applying **multiscale filtering** to the temporal histograms before performing depth reconstruction. The idea is to smooth the photon count cube `zcube`$[:, :, t]$—i.e., the image of photons detected at time t—using a mean filter (average kernel) of size $k \times k$.

This kernel is defined as:

$$K_k(i, j) = \frac{1}{k^2}, \quad \forall i, j \in \left\{ -\left\lfloor \frac{k}{2} \right\rfloor, \dots, \left\lfloor \frac{k}{2} \right\rfloor \right\}$$

This corresponds to a uniform low-pass filter, which replaces the value of each pixel by the average of its neighboring values within a local $k \times k$ window.

Mathematically, for each spatial coordinate (x, y) and each temporal bin t, the filtering operation gives a new value:

$$\hat{z}(x, y, t) = \sum_{i=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} \sum_{j=-\lfloor \frac{k}{2} \rfloor}^{\lfloor \frac{k}{2} \rfloor} K_k(i, j) \cdot z(x + i, y + j, t)$$

Let $z_t(x, y) = z_{\text{cube}}[x, y, t]$ denote the photon count at spatial location (x, y) and temporal bin t, and let $\tilde{z}_t(x, y)$ denote the filtered result after applying the spatial smoothing kernel.

We define

$$z_t(x, y) = z_{\text{cube}}[x, y, t],$$

and let

$$\tilde{z}_t(x, y)$$

denote the filtered result.

This filtering is applied for each temporal bin, resulting in a total of T "2D convolutions" for each kernel size k.

In practice, we invoke the `scipy.ndimage.convolve` function, performing the convolution in `'reflect'` mode. This mode applies symmetric padding along the image edges: at pixels near the boundary, out-of-bounds values are approximated by reflection. This avoids artifacts that would otherwise arise from zero or constant padding, thereby maintaining uniform filtering across the entire image domain, including edges.

From a processing standpoint, this corresponds to independent spatial smoothing for each temporal image slice. The goal is to reduce random variation (noise) in the photon histograms. This approach is based on the assumption that, within a small spatial neighborhood, adjacent pixels share common depths and, hence, exhibit similar histogram shapes. The filtering leverages this spatial redundancy to remove noise without significantly distorting the signal of interest (e.g., the Gaussian peak associated with time-of-flight).

The benefit of multiscale filtering is the ability to employ several kernel sizes $k \in \{1, 3, 5, 7, 9\}$, each representing a different trade-off:

- $k = 1$: no filtering, raw histograms (high sensitivity to noise),
- $k = 3$ or 5: light to moderate filtering, balancing noise attenuation and detail preservation,
- $k \geq 7$: strong filtering, better noise suppression with potential loss of fine depth structures.

Each scale produces an independent version of the filtered histograms, which are reconstructed separately. By analyzing these reconstructions (e.g., using metrics like RMSE), we can investigate how neighborhood size impacts spatial smoothing effectiveness.

After multiscale processing, depth reconstruction is further enhanced by statistically fusing the outputs obtained at different kernel sizes. This fusion is crucial for noise reduction, stable pixel-wise depth estimation, and improved spatial coherence.

Two statistical fusion strategies were explored:

- **Median fusion:** prioritizes robustness to outliers,
- **Mean fusion:** favors structural smoothing and averaging.

C – Median Fusion Strategy:

The **median** is the middle value of a sorted list. More precisely, for a set of K values, the median is the value that lies at the center of the sorted sequence. If K is odd, the median is the exact middle value; if K is even, it is defined as the average of the two middle values.

In contrast to the mean, the median is *not* influenced by outliers or extreme values, making it a noise-robust estimator.

1. **Histogram Fusion Using Median** Each pixel's temporal response is captured for different kernel sizes, resulting in multiple histograms $z_1, z_2, \ldots, z_K$ for each pixel. Each histogram is of length T, corresponding to the number of time bins.

   To obtain a single noise-robust histogram per pixel, we compute:

   $$z_{\text{median}} = \text{median}(z_1, z_2, \ldots, z_K)$$

   This approach ensures that any unusual peaks present in only a few scales are disregarded, preserving only the central tendency across the set of histograms.

2. **Depth Map Fusion Using Median** The same concept is applied to the reconstructed depth maps $D_1, D_2, \ldots, D_K$. For each pixel location (i, j), the median value across all scales is computed as:

   $$D_{\text{median}}(i,j) = \text{median}(D_1(i,j), D_2(i,j), \ldots, D_K(i,j))$$

   This results in a depth map that is more robust to scale-specific artifacts or anomalies, producing a representation that is less sensitive to the particular choice of kernel size.

D – Mean Fusion Strategy:

The mean, or arithmetic average, is calculated by adding up all the values and dividing by their count:

$$\text{mean}(x_1, \ldots, x_K) = \frac{1}{K} \sum_{k=1}^{K} x_k$$

It provides a central value measure but is responsive to outliers. If a number is quite higher or lower compared to the rest, it can change the average significantly.

1. **Histogram Fusion Using Mean** In order to get a representative histogram from the multiscale responses, the average histogram is calculated for every pixel:

$$z_{\text{avg}} = \frac{1}{K} \sum_{k=1}^{K} z_k$$

This method smooths out personal variations and brings out common patterns at different scales. But strong outliers could still influence the outcome.

2. **Depth Map Fusion Using Mean** In the same way, pixel-wise mean is computed for the reconstructed depth maps:

$$D_{\text{avg}}(i,j) = \frac{1}{K} \sum_{k=1}^{K} D_k(i,j)$$

4.2.3 Depth estimation using MF and ARGMAX Filtering

Post the multiscale processing stage, depth map reconstruction involves calculating one depth value per pixel from the data obtained. Two approaches were employed to implement this operation: a straightforward approach utilizing the argmax operator, and a more robust approach utilizing a matched filter.

A – Depth Reconstruction Using Argmax:

The first method relies on a straightforward strategy. For each pixel, the maximum value of the multiscale response vector is selected—that is, the most likely. Then, using the index of this maximum value, one goes to fetch the corresponding depth from the list of normalized depth values.

Mathematically, given: $\mathbf{z}_i = [z_{i1}, z_{i2}, \dots, z_{iN}]$ be the multiscale response vector for the i-th pixel, and $\mathbf{d} = [d_1, d_2, \dots, d_N]$ the corresponding vector of discrete depth values.

The estimated depth $\hat{d}_i$ is then given by:

$$\hat{d}_i = d_{k^*} \quad \text{where} \quad k^* = \arg\max_k z_{ik}$$

This technique has the advantage of being very fast and easy to apply. It provides a good starting point for an initial reconstruction. However, it suffers from some disadvantages, particularly in noisy conditions or if the response distribution is not well characterized (i.e., if there are several values close to the maximum). In this case, `argmax` will give unstable or erroneous depth estimates.

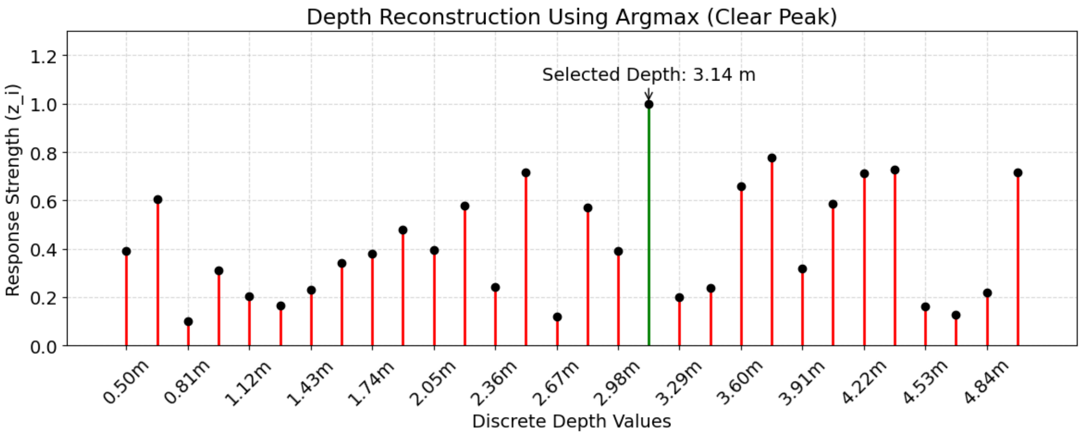


Figure 4.3: Illustration of the `argmax` method for depth reconstruction.

As illustrated in Figure 4.3, this method is based on selecting the maximum value from the response vector $\mathbf{z}_i$, and using its index $k^\star$ to retrieve the corresponding discrete depth $d_{k^\star}$.

B – Depth Reconstruction using Matched Filter :

To make depth estimation more robust, a secondary method based on a matched filter was incorporated. This involves convolving each pixel's response vector with a pre-defined filter that will highlight response shapes close to the expected (i.e., peaked and centered).

The matched filter is implemented as a discrete Gaussian window centered at the midpoint of the response vector. Let N be the number of discrete depth values, and $c = \left\lfloor \frac{N}{2} \right\rfloor$ the center index. The matched filter is computed as:

$$h[k] = \exp\left(-\frac{(k-c)^2}{2\sigma^2}\right)$$

The filter is then normalized so that the sum of all its values equals 1:

$$h[k] \leftarrow \frac{h[k]}{\sum_{j=0}^{N-1} h[j]}$$

For each pixel i, let its response vector be $\mathbf{z}_i = [z_{i0}, z_{i1}, \ldots, z_{i,N-1}]$. The filtered response is computed by a 1D convolution in "same" mode (to keep the length), which can be written as:

$$s_i[k] = (\mathbf{z}_i * h)[k]$$

Finally, the predicted depth is the location where the filtered response is maximum:

$$\text{where} \quad k^* = \arg\max_k s_i[k]$$

The actual depth value is then read from the normalized depth vector:

$$\hat{d}_i = d_{k^*}$$

This matched filter method enhances the depth map reconstruction accuracy and stability, particularly when raw responses are noisy or ambiguous. The width of the Gaussian is controlled by the parameter σ (named `sigma_filter` in the code) and can be adjusted to fit the expected signal spread.

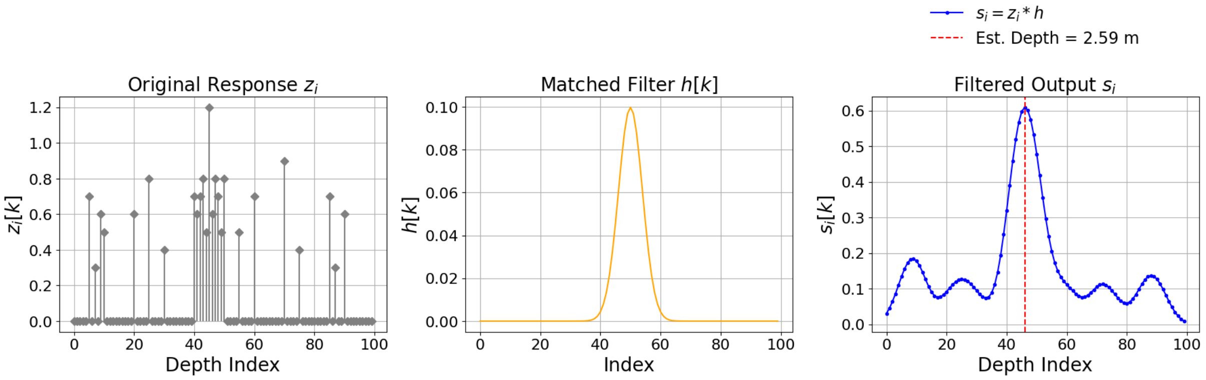


Figure 4.4: Matched filter depth estimation. Left: original response $\mathbf{z}_i$. Center: Gaussian filter $h[k]$. Right: filtered output $\mathbf{s}_i = \mathbf{z}_i * h$ with estimated depth.

Figure 4.4 illustrates the matched filter approach for depth estimation. The first panel shows the original response vector $\mathbf{z}_i$, where the peak is not very sharp. The second panel displays the Gaussian-shaped matched filter $h[k]$, centered around the expected location. In the third panel, the filtered response $\mathbf{s}_i = \mathbf{z}_i * h$ is shown, where the convolution results in a smoother and more stable peak. The estimated depth corresponds to the index of the maximum filtered response, which is then mapped back to the discrete depth value.

4.2.4 Plug-and-Play strategy

To tackle the heterogeneity and complexity of real-world scenes, our method leverages a Plug-and-Play strategy inspired by state-of-the-art computational imaging and processing LiDAR data research [33]. This allows us to plug-in independently designed modules—each handling a specific subtask such as noise removal, upsampling, or saliency estimation—into one and the same pipeline.

Within such a paradigm, deep learning-based models (e.g., UC-Net for saliency prediction) can be supplemented with model-based algorithms (e.g., matched filtering or statistical fusion) that share both interpretability and data-adaptive characteristics. Modular design raises flexibility levels so that elements can be replaced or reconfigured based on specific input conditions, such as fog-deteriorated depth maps, sensor noise, or spatial displacement.

This also increases overall generalization to out-of-sample data and simplifies transfer of the framework to other tasks. The Plug-and-Play also facilitates real-time or near-real-time processing through optimization of each module in isolation with overall system performance maintained.

4.3 Conclusion

This chapter presented a comprehensive approach to improving the reconstruction of depth maps generated by dToF sensors, taking into account the noise typically present in real-world data. By simulating various conditions through the PPP (Photons Per Pixel) and SBR (Signal-to-Background Ratio) parameters, and applying techniques such as multiscale processing, statistical fusion, and reconstruction filters (argmax and matched filter), we enhanced the robustness and accuracy of the DVSR and HVSR algorithms. The use of objective metrics such as RMSE (Root Mean Square Error) and AE (Absolute Error) enabled a rigorous quantification of the improvements achieved at each stage of the proposed pipeline. The results obtained for each component of the approach will be presented and analyzed in the following chapter.

CHAPTER 5 :

Experimental Results and Analysis

In this chapter, we present and analyze the results of our experiments. Our main objective was to simulate LiDAR data under real-world conditions by introducing noise and applying multiscale processing, as described in the previous chapter. To assess the accuracy of our approach, we use the Root Mean Square Error (RMSE) to compare the reconstructed outputs against the ground truth across the three different types of datasets.

Finally, as part of our high-level computer vision objectives, we present the results of the saliency detection process applied after all reconstruction and fusion steps.

5.1 Development Environment: Hardware and Software

To ensure high performance, reproducibility, and compatibility with the original DVSR (Consistent Direct Time-of-Flight Video Depth Super-Resolution) framework, a high-performance development environment was built. The environment supports intensive training and real-time inference on large-scale video depth data to standards compatible with CVPR 2023 research practice.

5.1.1 Hardware Configuration

A high-performance hardware configuration was required to handle high-resolution RGB sequences and low-resolution depth signals across thousands of frames. The specifications are as follows:

- **CPU:** Intel® Core™ i7 / AMD Ryzen™ 7 – High multi-threaded performance for data preprocessing and orchestration tasks
- **GPU:** NVIDIA RTX 3090 (24GB VRAM) – Large-scale training with CUDA Compute Capability $\geq$ 7.0 and mixed-precision acceleration
- **RAM:** 64 GB DDR4 – Enables smooth handling of video sequences for training and testing
- **Storage:** 1 TB NVMe SSD – Ensures quick access to datasets and intermediate results
- **Operating System:** Ubuntu 20.04 LTS / Windows 10 with WSL2 – Combines Linux-based flexibility with Windows compatibility

This setup ensures smooth experimentation with multi-view, spatio-temporal fusion architectures, and real-time inference pipelines.

5.1.2 Software Stack

The software environment is based on Python and the PyTorch + OpenMMLab ecosystem, bringing together best-in-class tools for training, video processing, and evaluation:

- **Language:** Python 3.8
- **Framework:** PyTorch 1.12.1 + CUDA 11.3
- **Vision/Audio Libraries:** `torchvision==0.13.1`, `torchaudio==0.12.1`
- **OpenMMLab Modules:**
  - `mmcv-full==1.7.0` – Core training infrastructure
  - `MMEditing` – Backbone framework for video super-resolution
  - `MMSegmentation==0.29.1` – Semantic guidance support
- **Utility Libraries:** `scipy`, `pyyaml`, `natsort`, `terminaltables`, `tqdm`

5.1.3 Environment Management

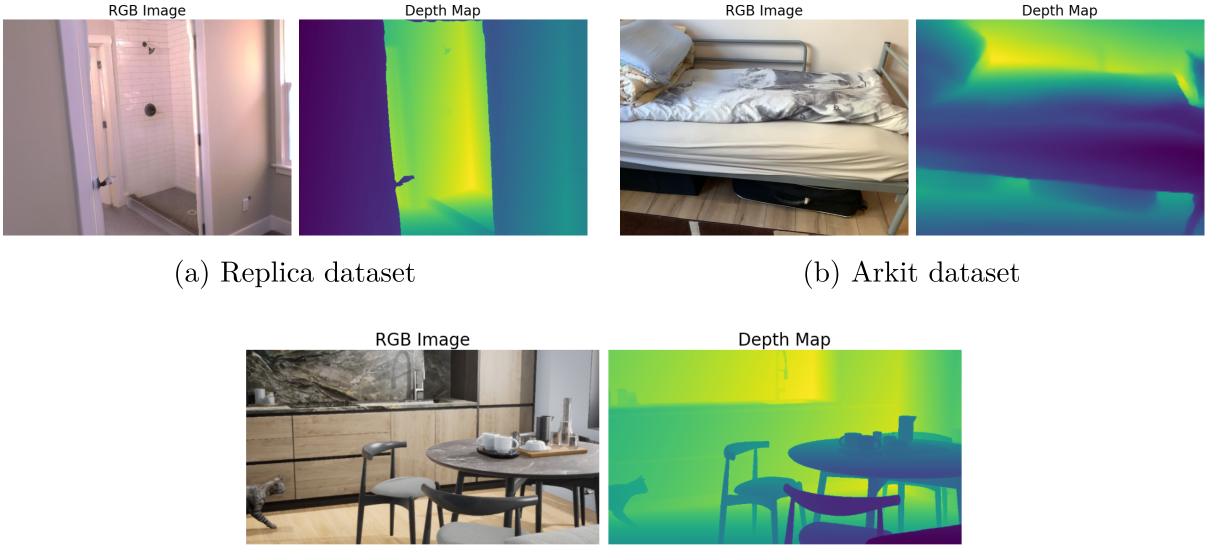
To foster reproducibility and avoid dependency conflicts, the entire project was encapsulated in a dedicated Conda environment:

```
conda create -name dvsr python=3.8
```

This isolated environment ensured a clean working space, full compatibility with CUDA-supported libraries, and stability during all experimental runs.

5.2 Dataset Overview

The datasets used in this work consist of three main sources: two synthetic datasets (Replica and DyDToF) and one real-world dataset (ARKit). Each serves a specific purpose in the training and evaluation of the DVSR and HVSR models.
Examples of the datasets are shown in Figure 5.1, where RGB images and their associated depth maps illustrate the diversity in content and resolution.



(a) Replica dataset

(b) Arkit dataset

(c) Dydtof dataset

Figure 5.1: Examples of the three datasets used in this study: (a) Replica, (b) Arkit, and (c) Dydtof

Table 5.1 summarizes the key properties and roles of the three datasets used, complementing the visual examples in Figure 5.1.

Dataset	Description	Characteristics	Purpose
Replica 5.1a	Synthetic RGB-D dataset generated from the Replica 3D indoor scenes using an image formation model.	- 100 RGB images + 100 depth maps - Resolution: 640×480 - Highly realistic textures and geometry	Ideal for depth reconstruction under static, controlled conditions.
ARKit 5.1b Apple's Augmented Reality Kit	Real-world dataset captured via iPhone with ARKit and ToF sensor.	- 50 RGB + 50 depth images - Resolution: 256×192 (downsampled) - Contains real-world noise and artifacts	Used to evaluate generalization of models to real sensor data.
DyDToF 5.1c	Synthetic dataset created using Unreal Engine to emulate direct ToF sensors.	- 100 RGB images + 100 depth maps - Resolution: 640×352 - High visual and structural realism	Designed for training/testing on temporally coherent, dynamic, synthetic ToF data.

Table 5.1: Comparison of the Replica, DyDToF, and ARKit datasets used in our experiments.

5.3 Performance Metrics Used for Evaluation

To measure the quality of the reconstruction obtained with the different methods used, we used these two common metrics :

- **RMSE**:

  The root mean square error (RMSE) measures the average difference between a statistical model's predicted values $\hat{x}_i$ and the actual values x_i :

  $$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

  It is outlier-sensitive, with very large errors being harshly penalized by squaring. Thus, it is a useful measure to use for the detection of reconstructions with very large local errors.

- **AE** :

  The absolute error (AE) measures the absolute difference between observed $\hat{x}_i$ and original values x_i :

  $$\text{AE}_i = |x_i - \hat{x}_i|$$

  It can be averaged for all the samples to obtain the MAE (Mean Absolute Error). Unlike RMSE, AE is not as sensitive to single large errors and provides a more stable and representative overall error estimate.

5.4 Results of synthesising noisy data :

Figure 5.2 presents example results from the three datasets—Replica, DyDToF, and ARKit—using two reconstruction methods: Matched Filter and Argmax.

The figure is split into two sections: the left side shows results from the Matched Filter, while the right side displays those from the Argmax method.

Each row corresponds to a different (PPP, SBR) pair, highlighting the performance of both methods under varying conditions across all datasets. For each pair, the corresponding RMSE values—computed with respect to the ground truth—are also displayed, allowing for a quantitative comparison of the reconstruction accuracy.

The results are analyzed and discussed in the following section.

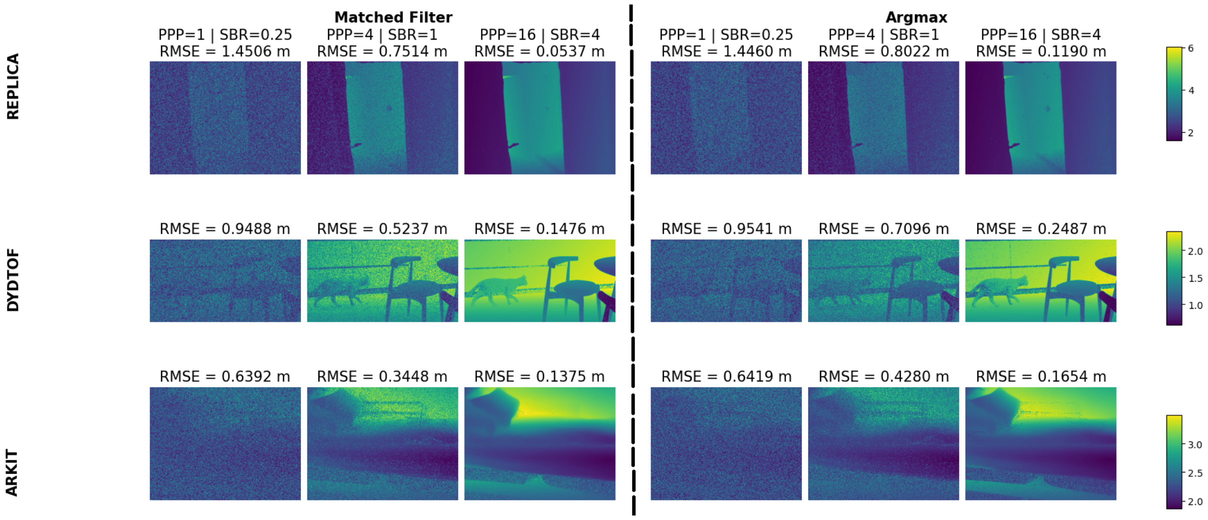


Figure 5.2: Results of synthesising noisy data using the two methods : Matched filter and Argmax

5.4.1 Using Matched filter for the reconstruction :

In this case, a matched filter was used to reconstruct the depth map.

At low values of PPP (Photons Per Pixel) and SBR (Signal-to-Background Ratio), the RMSE is significantly high—1.45 m for the DVSR dataset and 0.94 m for the DYDTOF dataset—indicating substantial reconstruction error.

For medium PPP and SBR values, the depth maps appear less noisy, and more scene details are visible. The RMSE in this case is moderate; for example, the DVSR dataset shows an RMSE of 0.75 m, which is the highest among the three datasets at this level.

When the PPP is high, the depth maps are visually accurate and much cleaner, although some fine details may still be lost. Overall, the reconstruction quality improves considerably as PPP and SBR increase.

5.4.2 Using Argmax :

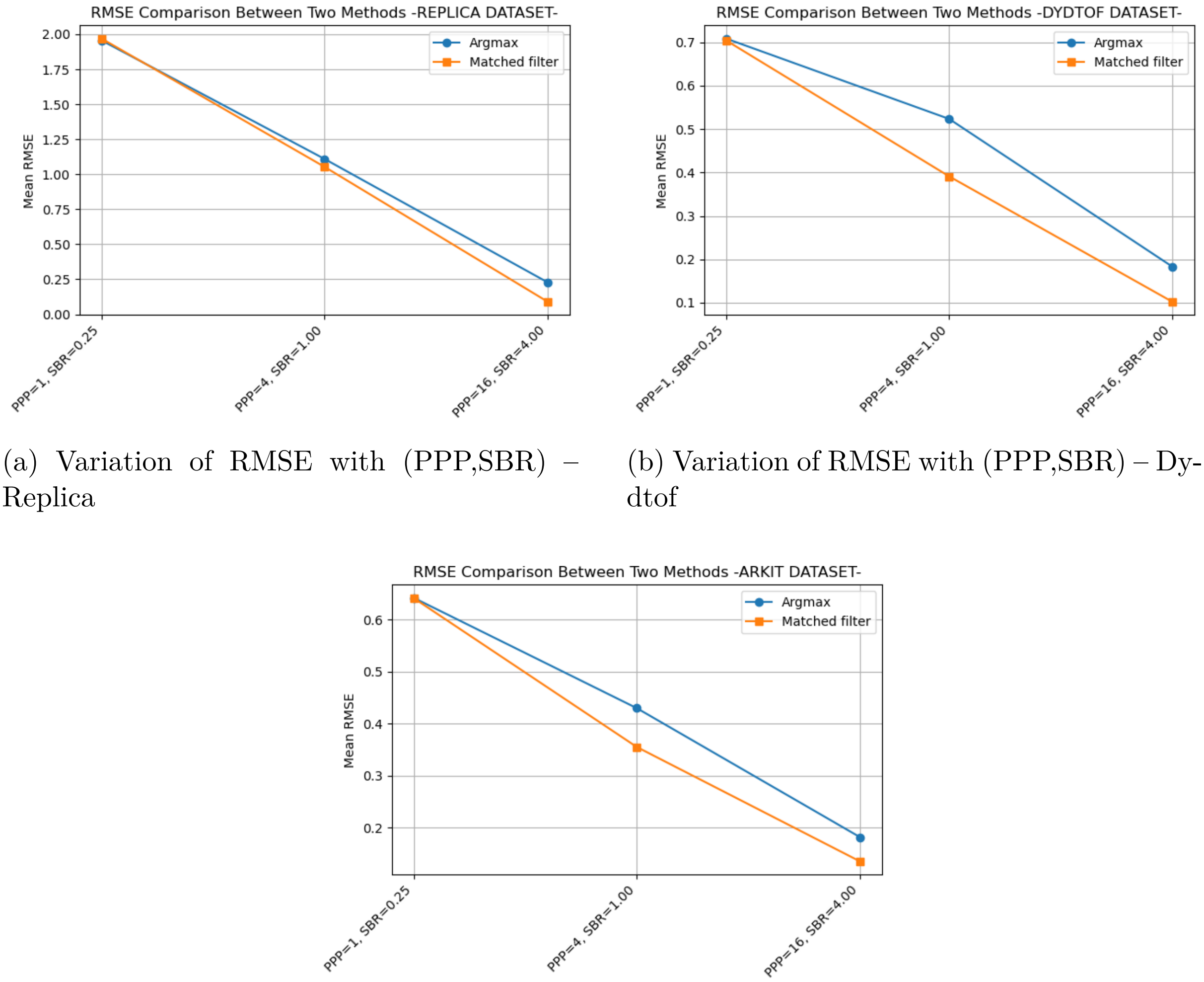
In this case, a naive reconstruction was used :

As we did for the matched filter, we evaluated the reconstruction performance using the *argmax* method under varying PPP and SBR conditions. At low values of PPP and SBR, the RMSE is significantly high—1.45 m for the DVSR dataset and 0.95 m for the DYDTOF dataset—indicating substantial reconstruction error and noisy depth maps.

For medium PPP and SBR values, the depth maps become less noisy, and more scene details are visible. The RMSE is moderate; for instance, the DYDTOF dataset shows an RMSE of 0.7 m. At high PPP and SBR, the depth maps are much cleaner and visually accurate, though some fine details may still be missed. Overall, as with the matched filter, the reconstruction quality improves significantly with increasing photon counts and signal quality.

5.4.3 Comparison between the two methods of reconstruction :

We compare the reconstruction performance of the Matched Filter and Argmax methods across the three datasets under varying PPP and SBR conditions. For this analysis, we calculated the mean RMSE over all frames for each dataset and plotted the corresponding graphs.
The following figure 5.3 presents the graphs.



(a) Variation of RMSE with (PPP,SBR) – Replica

(b) Variation of RMSE with (PPP,SBR) – Dydtof

(c) Variation of RMSE with (PPP,SBR) – Arkit

Figure 5.3: RMSE as a function of PPP and SBR for the three datasets: a- Replica, b- Dydtof, and c- Arkit.

These are some comments about the graphs 5.3 :

- The Matched Filter provides more robust reconstruction, particularly in scenarios with higher Photon Per Pixel (PPP) and Signal-to-Background Ratio (SBR).
- The performance gap widens as the quality of the data improves, confirming the superior denoising and localization capability of the Matched Filter compared to the simpler Argmax approach.

- At low data quality ($\text{PPP} = 1, \text{SBR} = 0$), both methods are similarly affected, with no significant distinction observed.

Table 5.2 compares Argmax and Matched Filter in terms of noise sensitivity, accuracy, and robustness under low photon conditions, highlighting their respective strengths.

Method	**Noise impact**	**Accuracy (RMSE)**	**Robustness (Low PPP,SBR)**	**Main Advantages**
Argmax	High	Lower (variable)	Fails under low photons	Simple and fast to compute; suitable for clean data
Matched Filter	Low	Higher (stable)	Robust to noise + sparse signals	Improved accuracy under noise; better depth localization; suitable for real-world degraded data

Table 5.2: Comparison between Argmax and Matched Filter reconstruction methods under varying photon conditions.

In summary, the Matched Filter consistently outperforms the Argmax method in terms of reconstruction accuracy, noise robustness, and adaptability to low-photon conditions. Its superior performance across diverse datasets and varying PPP/SBR configurations demonstrates its effectiveness in handling challenging scenarios typical of real-world data. Given these advantages, we adopt the Matched Filter as the primary reconstruction method for depth map generation in the subsequent stages of our processing pipeline.

5.5 Results of the multiscale :

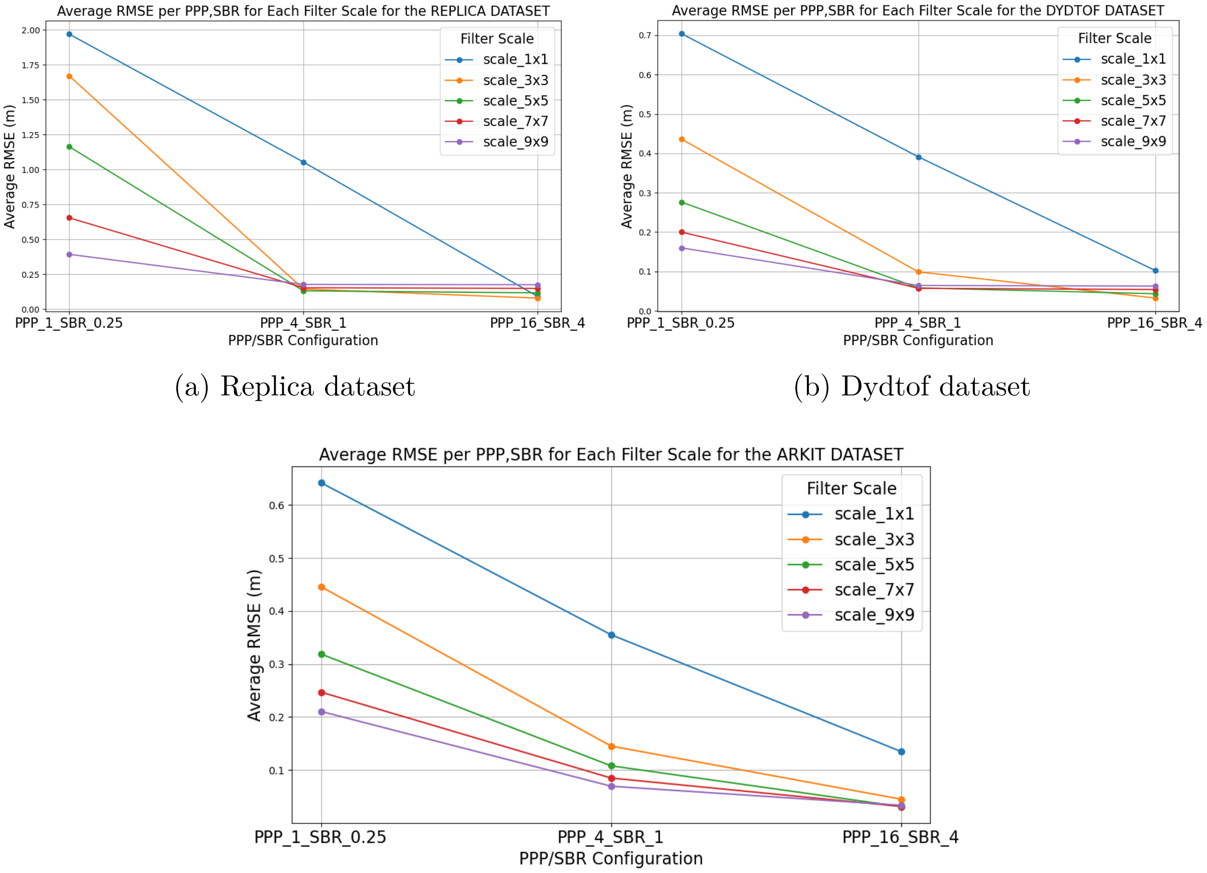
As demonstrated earlier, the matched filter performs better, so it will be used in the following steps.

Figure 5.4 illustrates the average RMSE curves across different scales under varying values of PPP and SBR for the three datasets.

As expected, RMSE generally decreases with increasing PPP and SBR values, since higher values imply more reliable and informative input data, resulting in reconstructions that are closer to the ground truth. Additionally, we observe that increasing the kernel size systematically reduces RMSE across all datasets, particularly in scenarios with low PPP. For instance, in the Replica dataset, the RMSE drops significantly from approximately 2 m at scale 1 to 0.4 m at scale 9. This improvement is largely attributed to noise reduction, as larger kernels capture more neighborhood context, enhancing the robustness of the depth estimation in highly noisy cases.

However, this trend reverses for medium and high PPP settings. Starting from scale 3, increasing the kernel size leads to a noticeable rise in RMSE. This occurs because averaging over a larger neighborhood modifies depth values that are already accurate, resulting in the loss of meaningful spatial details. In such cases, overly large kernels blur important features and degrade the quality of the reconstruction.

Therefore, increasing the kernel size introduces a trade-off: while it reduces noise, it also leads to a loss of spatial resolution. To address this, we employ a multiscale approach using kernel sizes of 1, 3, 5, 7, and 9, and fuse the results using both mean and median strategies based on depth map and histogrames. This enables a balance between preserving fine details and suppressing noise, depending on the characteristics of the input data.



(a) Replica dataset

(b) Dydtof dataset

(c) Arkit dataset

Figure 5.4: Average RMSE per PPP,SBR for each scale for the 3 datasets

5.5.1 Depth-map-based method:

We first reconstructed the depth maps at all individual scales, and then calculated the mean and median of the resulting depth values.

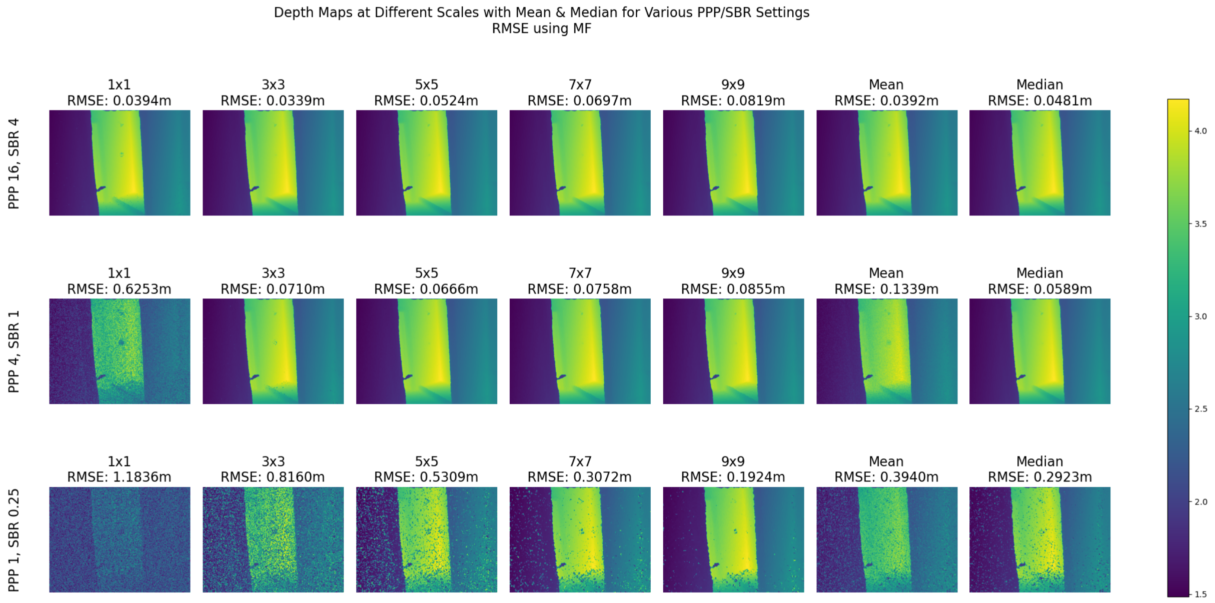


Figure 5.5: Depth maps at different scales with Mean and Median results

Figure 5.5 illustrates the results of the multiscale implementation using various kernel sizes on the DVSR (Replica) dataset across different (PPP, SBR) configurations.

Each row corresponds to a specific (PPP, SBR) pair, while each column shows the reconstruction results for a given kernel size (1, 3, 5, 7, and 9). The final two columns display the results obtained using the mean and median of the multiscale outputs.

Below each reconstructed depth map, the corresponding RMSE value, computed with respect to the ground truth, is provided to facilitate quantitative evaluation.

Visually, we observe that in the most challenging scenario (PPP = 0.25, SBR = 1), increasing the kernel size significantly reduces noise. A similar denoising effect is also evident in the case of (PPP = 4, SBR = 1), where larger kernels lead to noticeably cleaner reconstructions.

The mean and median across the five scales yield significantly improved results compared to the original noisy data, demonstrating the effectiveness of multiscale aggregation in enhancing reconstruction quality. Below are some observations regarding the RMSE behavior:

- For low PPP and SBR values, increasing the kernel size results in a notable decrease in RMSE, dropping from 1.18m to 0.19m. Furthermore, the mean and median aggregations outperform the original noisy input, confirming their usefulness in low-quality data scenarios.
- For medium PPP and SBR settings, the RMSE decreases progressively up to the 7×7 scale, after which a slight increase of around 10cm is observed. This suggests an optimal kernel size beyond which further smoothing may degrade the quality.

- For high PPP and SBR configurations, the original data is already of high quality. In this case, applying multiscale smoothing tends to increase the RMSE across all kernel sizes. However, the aggregated mean result remains acceptable and may still be retained depending on the application needs.

5.5.2 Histogram-based method: :

For each reconstructed depth map, we constructed a cube of histogram values corresponding to the different kernel sizes (scales). This cube represents the temporal histogram distribution at each pixel across the five selected scales (1, 3, 5, 7, and 9). From this 3D structure, we computed the mean and median histograms at each pixel location to generate aggregated reconstructions that benefit from multiscale smoothing while preserving useful spatial features.

Figure 5.6 illustrates the cube histogram for the specific case of PPP = 16 and SBR = 4 at diffrent scales, which corresponds to a high-quality data scenario

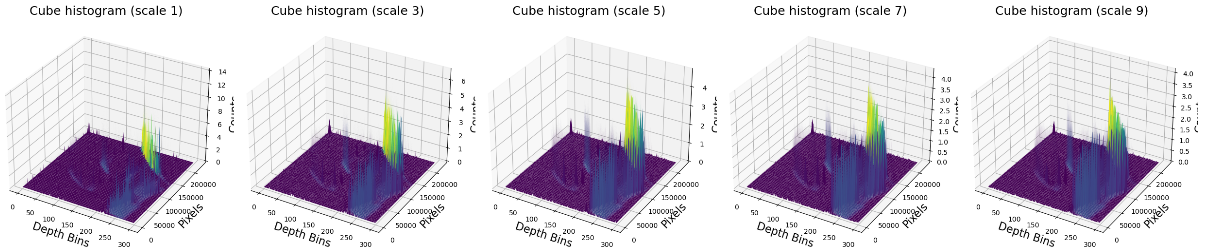


Figure 5.6: Cube histograms of the scales for PPP = 1 | SBR =0.25

.

Since the cube histogram does not reveal the individual histograms of each pixel, we selected a random pixel—located at coordinates (400, 200)—along with its 3×3 neighborhood which are in figure 5.7.

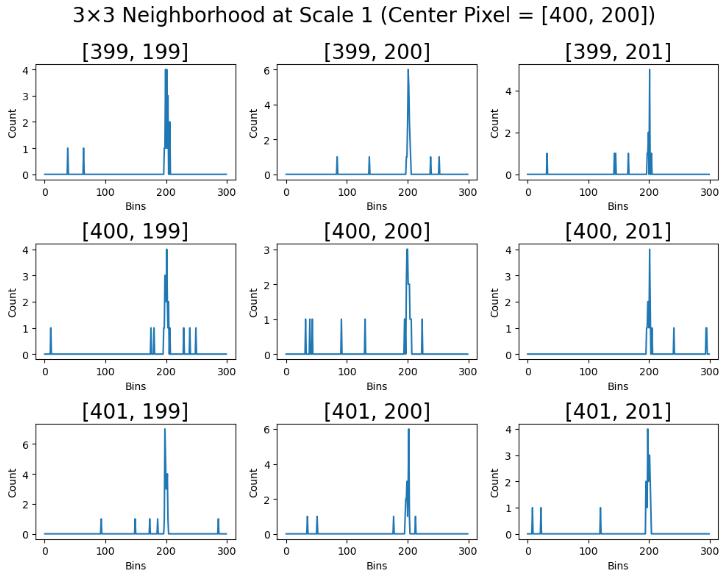


Figure 5.7: 3x3 neighborhood at scale 1 for the pixel (400,200)

Figure 5.8 presents the histograms corresponding to the selected pixel across multiple scales, highlighting how the distribution evolves with scale variation.

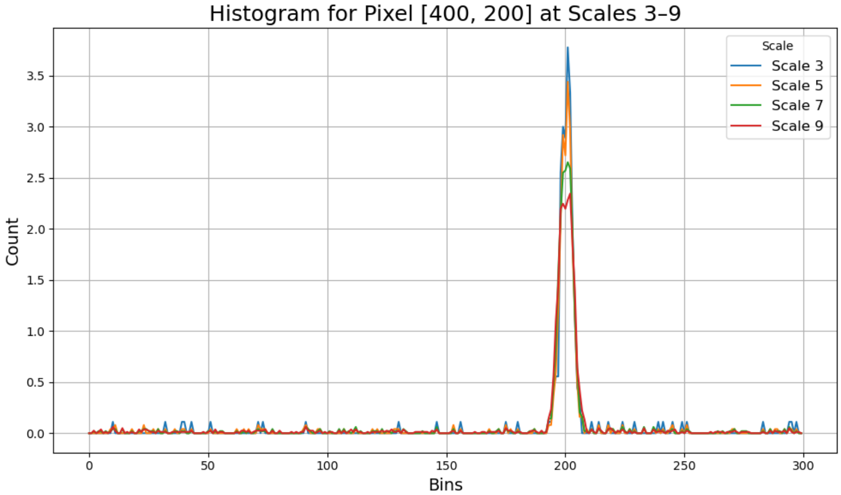


Figure 5.8: Histograms of the selected pixel (400,200) at diffrent scales

Subsequently, we applied both mean and median fusion across all scales. The results for the selected pixel are shown in Figure 5.9, while the aggregated cube histogram after multiscale fusion over the entire image is depicted in Figure 5.10. What follows is a comparison of the obtained results.

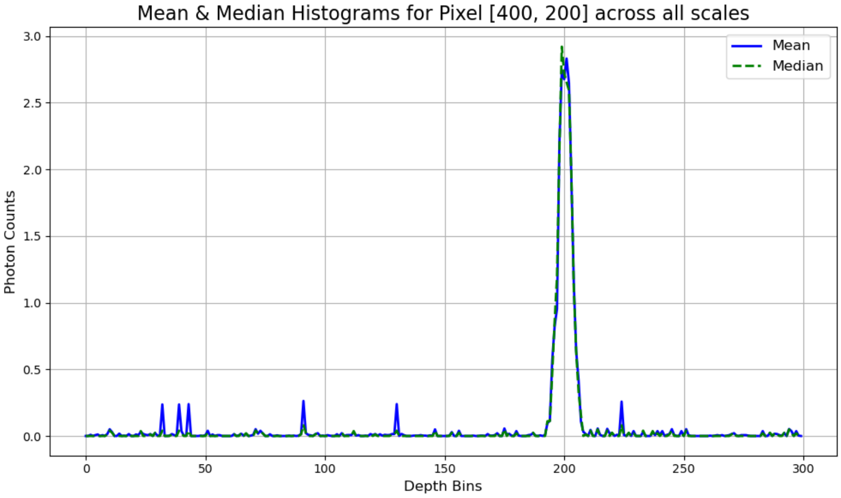


Figure 5.9: histograms of the mean and median results for the selected pixel (400,200) for PPP = 16 | SBR =4

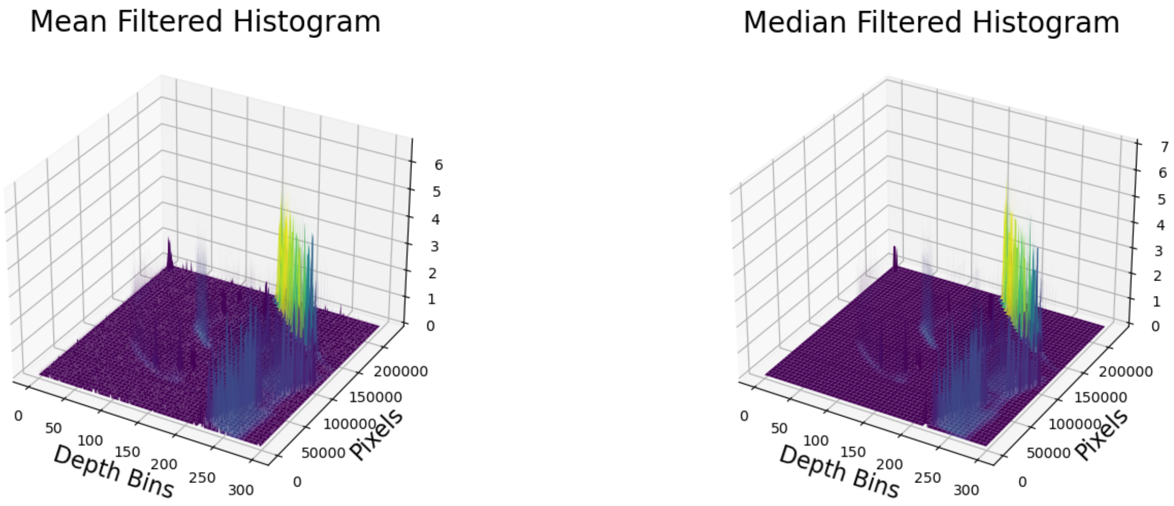


Figure 5.10: Cube histograms of the mean and median results for PPP = 16 | SBR =4

5.5.3 Comparison between the two methods :

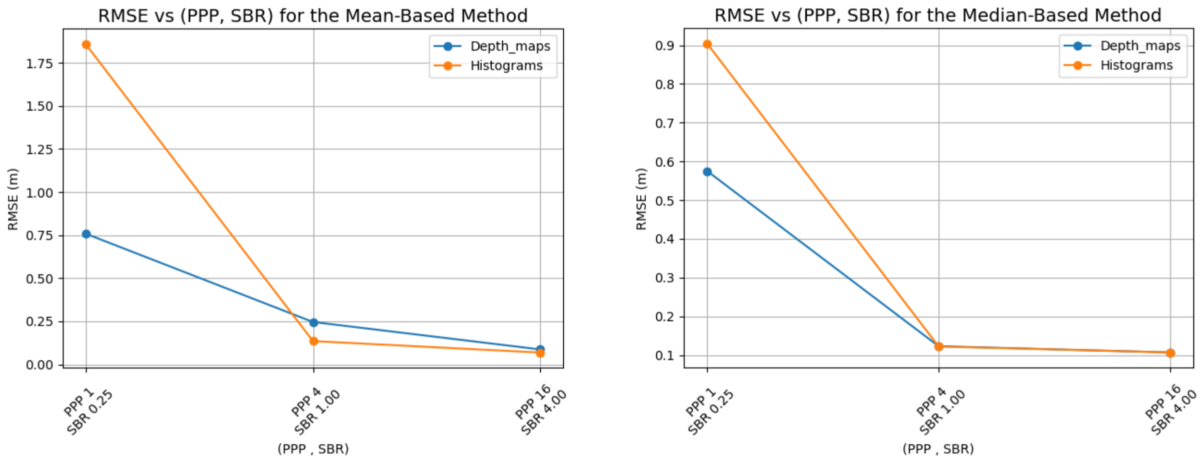


Figure 5.11: Comparison of RMSE as a function of PPP and SBR for the mean-based (left) and median-based (right) fusion strategies.

To compare the effectiveness of multiscale fusion strategies, we plotted the RMSE for different PPP and SBR combinations using two aggregation techniques: the **mean** and the **median**. The results are shown separately for depth maps and histograms 5.11.

The following key points summarize the observations from the two plots:

- **Mean-Based Fusion (Left Graph)**:
  - In the most challenging setting (PPP = 1, SBR = 0.25), depth maps provide significantly better results than histograms, with RMSE around 0.75 m versus 1.85 m.
  - As the data quality improves (higher PPP and SBR), the RMSE decreases noticeably for both methods.

  - ○ For high-quality data (PPP = 16, SBR = 4), histograms slightly outperform depth maps, both achieving RMSE values below 0.1 m.

- **Median-Based Fusion (Right Graph)**:
  - ○ Median fusion shows greater robustness to noise, particularly in low-quality scenarios.
  - ○ At PPP = 1, SBR = 0.25, it yields lower RMSE than the mean approach for both depth maps and histograms.
  - ○ As PPP and SBR increase, the gap between depth maps and histograms narrows, with both methods reaching similar low RMSE values.

- **Overall Findings**:
  - ○ Multiscale fusion significantly enhances reconstruction quality regardless of the method.
  - ○ Depth maps tend to be more reliable in noisy settings.
  - ○ Histogram-based methods become more effective when the signal-to-background ratio and PPP increase.
  - ○ Median fusion is generally more robust than the mean strategy across all conditions.

5.6 DVSR and HVSR results

Before we proceed to explicit quantitative and qualitative results, we present a short overview of the evaluation methodology. Having generated temporal response data for various reconstruction methods—namely, PPP (Photon per-Pixel) and SBR (Signal to Background Ratio)—on diverse spatial scales using our multiscale processing pipeline, we performed statistical fusion with both mean and median in order to increase robustness. The fused results were then processed with two recent state-of-the-art reconstruction algorithms:

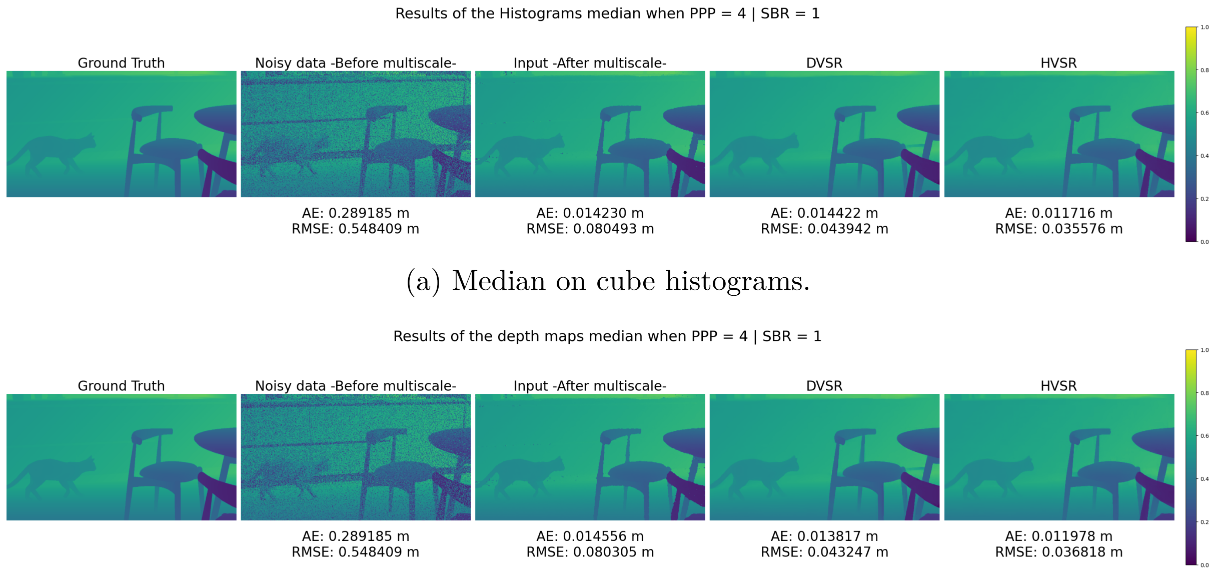
- **DVSR** (Depth Video Super Resolution).
- **HVSR** (Histogram Viedo Super Resolution).

These methods test the reliability and fidelity of the final reconstructed depth through both visual quality and performance measures such as RMSE and AE. In this section we present a detailed comparison of their performance for various types of fusion and qualities of input. In particular, we focus here on the medium-quality scenario, corresponding to PPP = 4 and SBR = 1.

5.6.1 Median based method

Figure 5.12 displays the DVSR and HVSR results generated using the median-based fusion method on the DYDTOF dataset.

Visually, combining multiscale aggregation with the super-resolution model yields significant enhancements, resulting in outputs that are sharper and more closely match the ground truth.



(a) Median on cube histograms.

(b) Median of depth maps

Figure 5.12: Comparison of the DVSR and HVSR results for the median method from cube histograms and depth maps for PPP = 4 , SBR = 1, on diffrent datasets

Comments :

- The multiscale preprocessing step visually enhances the data quality before feeding it into the model, resulting in improved performance when the model is applied.
- The median fusion strategy yields strong results. For instance, on the DYDTOF dataset, DVSR algorithm achieved AE = 0.0138m and RMSE = 0.0432m for the fused depth maps, and AE = 0.0144m and RMSE = 0.0439m for histogram fusion. This marks a substantial improvement compared to the noisy input data (AE = 0.289m, RMSE = 0.548m), showing much closer alignment with the ground truth
- At this stage, the depth map fusion method demonstrates superior performance compared to the histogram-based approach, even if the improvement in RMSE and AE is only by a few centimeters.

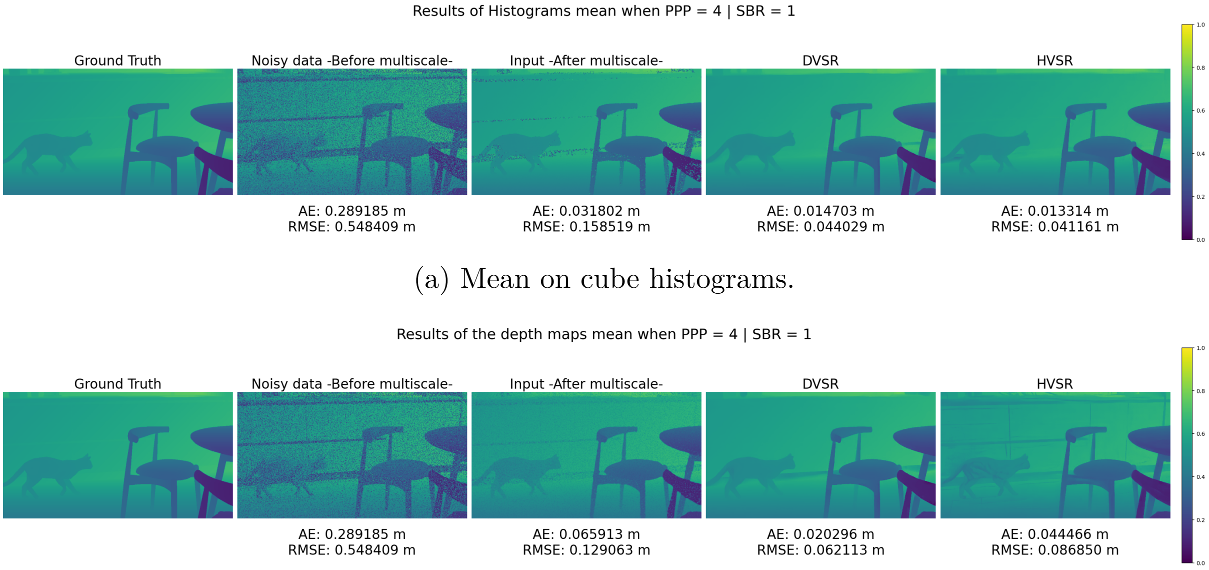
Table 5.3 presents the AE and RMSE values for the median-based methods applied to both histograms and depth maps from the ARKit dataset, which consists of real-world captures using an iPhone.A clear improvement is observed from left to right across the table, with both RMSE and AE progressively decreasing. Notably, median filtering on depth maps outperforms that on histograms. For instance, the AE decreases significantly from 0.1728m in the input to just 0.0134m after applying the multiscale approach combined with DVSR, highlighting the effectiveness of the method.

Config	**Input**		**Multiscale**		**DVSR**	
	AE	RMSE	AE	RMSE	AE	RMSE
Median - Histograms	0.1728	0.3516	0.0229	0.1043	0.0167	0.0594
Median - Depth maps	0.1728	0.3516	0.0164	0.0725	0.0134	0.0353

Table 5.3: AE and RMSE in meters for Median Methods

5.6.2 Mean based method

Figure below 5.13 presents the DVSR and HVSR results obtained using the mean-based fusion method. Visually, the use of multiscale aggregation combined with the super-resolution model leads to noticeable improvements, producing outputs that are clearer and more closely aligned with the ground truth.



(a) Mean on cube histograms.

(b) Mean of depth maps.

Figure 5.13: Comparison of the DVSR and HVSR results for the mean method from cube histograms and depth maps for PPP = 4 , SBR = 1, on diffrent datasets

Comments :

- The improvements are particularly noticeable for the DYDTOF dataset, which contains finer details. For example, in the case of the cat, the initial reconstruction was noisy, but after applying the method, the result becomes smoother while preserving the same depth information.
- Applying multiscale processing, which serves as a form of data pre-processing, enhances the quality of the data both visually and quantitatively, as reflected in improved RMSE and AE values.
- Examining the AE and RMSE values, the histogram-based mean fusion demonstrates superior performance compared to the depth map-based mean. For instance, in the DY-DTOF dataset, the HVSR method achieves an AE of 0.013m using histogram mean,

whereas the depth map mean yields an AE of 0.044m. This highlights the advantage of the histogram-based approach over the depth map method in this case.

Table 5.4 validates the results obtained for the DYDTOF dataset by presenting the AE and RMSE values for mean-based methods applied to both histograms and depth maps on the ARKit dataset. As evidenced by the table, the histogram-based approach consistently outperforms the depth map-based method.

Config	**Input**		**Multiscale**		**DVSR**	
	AE	RMSE	AE	RMSE	AE	RMSE
Mean - Histograms	0.1728	0.3516	0.0461	0.1646	0.0168	0.0593
Mean - Depth maps	0.1728	0.3516	0.0441	0.0911	0.0174	0.1668

Table 5.4: AE and RMSE in meters for Mean Methods

5.6.3 Comparison between the methods :

Figures 5.14 illustrate the variation of average RMSE as a function of (PPP, SBR) for the two output fusion strategies: median and mean, across different reconstruction methods (depth-dvsr, depth-hvsr, hist-dvsr, hist-hvsr).

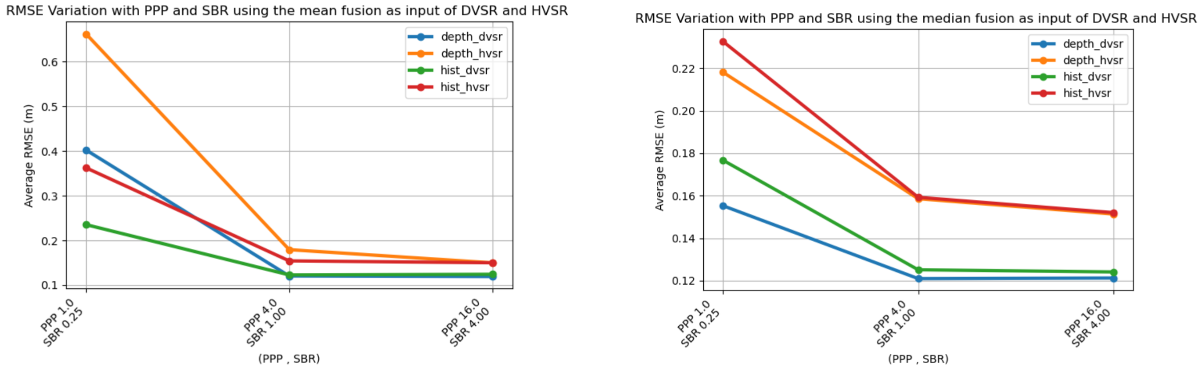


Figure 5.14: Comparison of RMSE as a function of PPP and SBR for the mean-based (left) and median-based (right) fusion strategies.

Median Fusion:

- The `depth_dvsr` method achieves the lowest RMSE values across most PPP and SBR combinations.
- Its performance improves significantly with increasing PPP and SBR, indicating high sensitivity to data quality and density.
- Histogram-based methods (`hist_dvsr`, `hist_hvsr`) show higher RMSEs, particularly under low SBR or PPP.

- The `hist_hvsr` method consistently yields the highest RMSE values, 0.24m for PPP = 1, SBR = 0.25

Conclusion: Median fusion is highly effective for depth-based methods, especially `depth_dvsr`, providing robust performance across various conditions.

Mean Fusion:

- The `depth_dvsr` method performs poorly under low PPP and SBR, with higher RMSEs than in the median case.
- Histogram-based methods (`hist_hvsr` and `hist_dvsr`) show more stable and competitive performance.
- The performance gap between histogram- and depth-based methods narrows.
- Mean fusion is more sensitive to outliers, which negatively affects depth-based reconstructions.

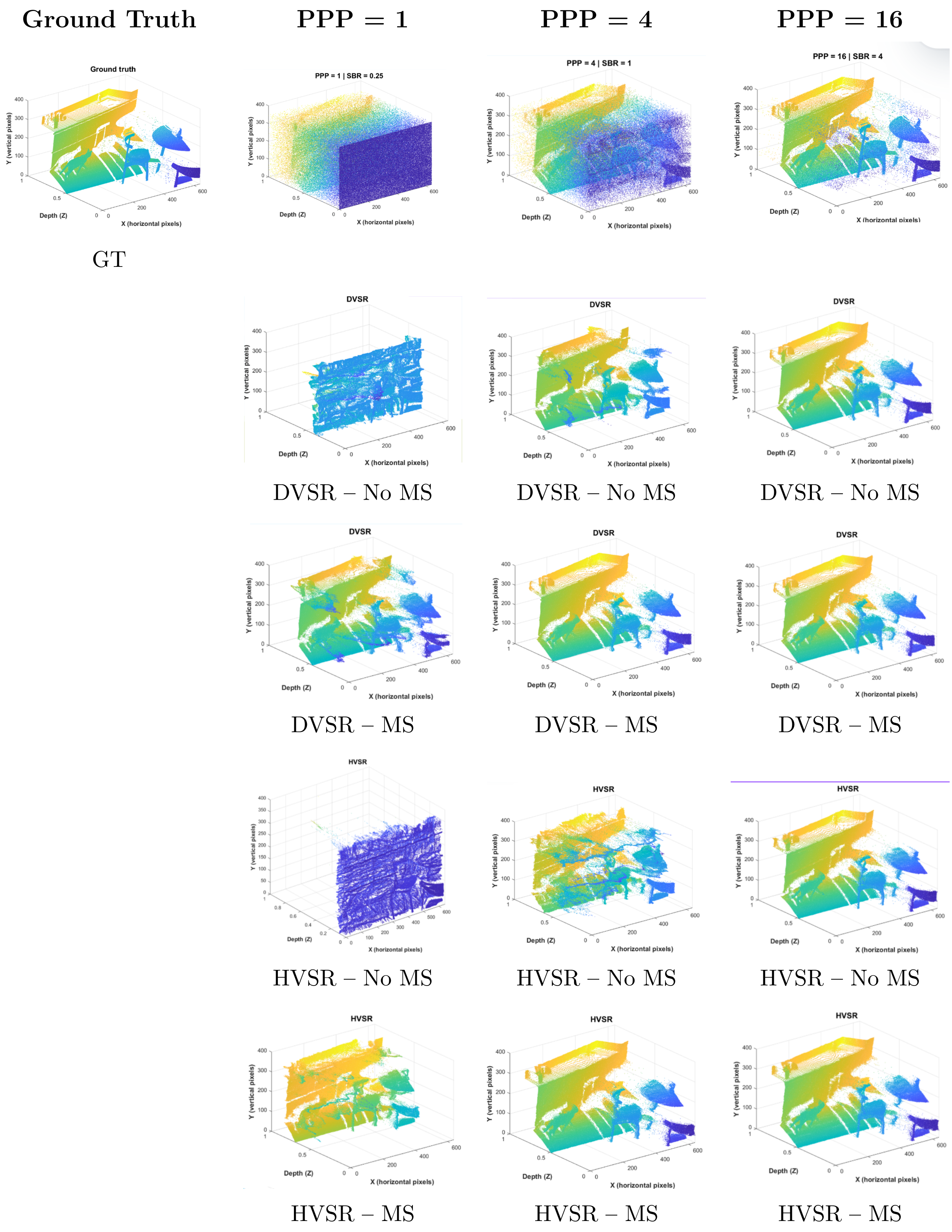
Conclusion: Mean fusion favors histogram-based methods due to their stability across varying input conditions.

Overall Insight: Median fusion provides superior robustness and accuracy for depth-based methods, while mean fusion may be preferable for histogram-based approaches in settings with variable or noisy inputs.

5.7 3D point cloud result visualization

Given the superior performance observed with mean fusion applied to histograms and median fusion applied to depth map, we opted to showcase the corresponding results in the subsequent analysis.

This figure 5.15visually reinforces the quantitative results presented earlier by showcasing median-filtered 3D point cloud reconstructions for both DVSR and HVSR methods across different PPP levels (1, 4, and 16), with and without multiscale filtering.



Ground Truth PPP = 1 PPP = 4 PPP = 16

GT

DVSR – No MS DVSR – No MS DVSR – No MS

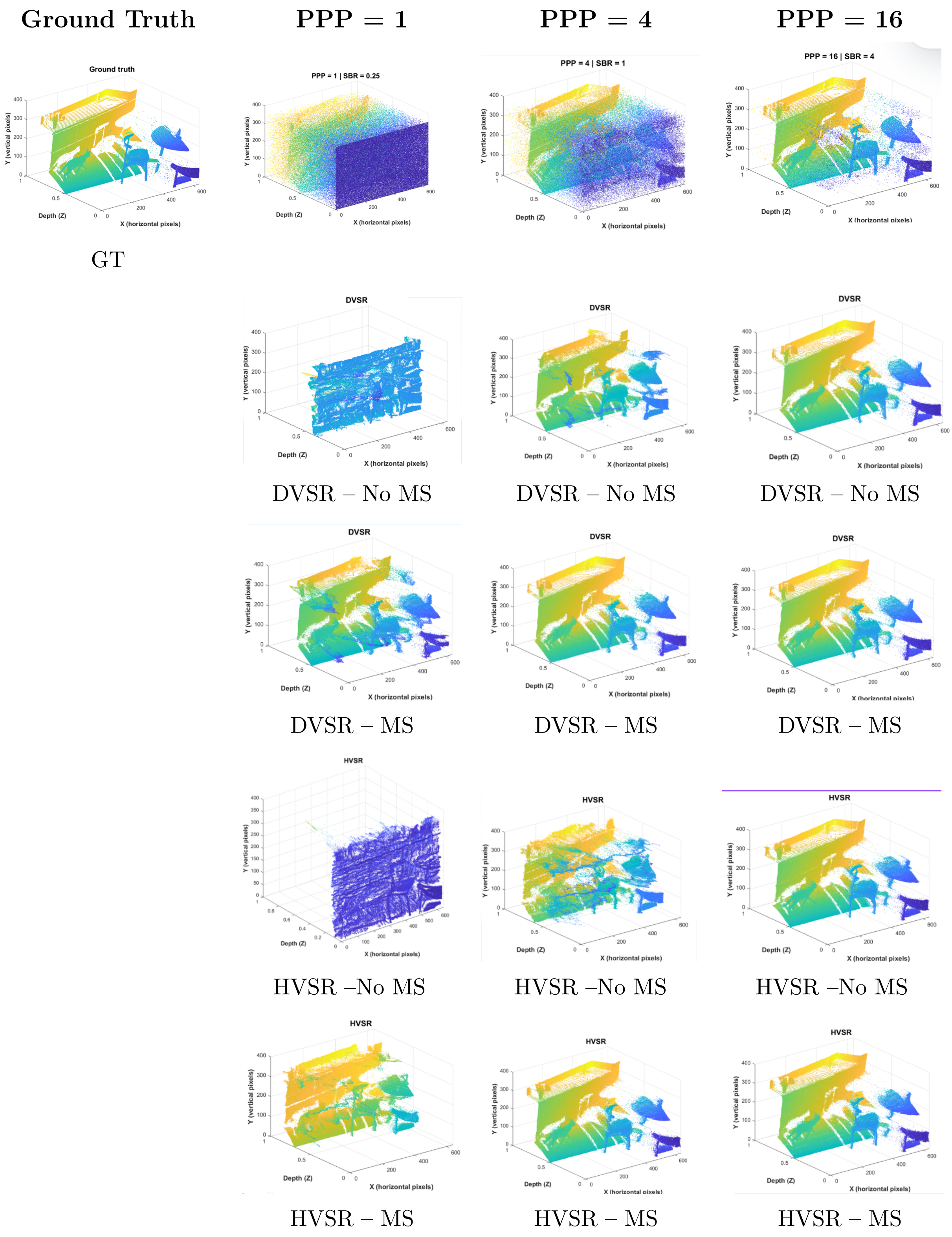
DVSR – MS DVSR – MS DVSR – MS

HVSR – No MS HVSR – No MS HVSR – No MS

HVSR – MS HVSR – MS HVSR – MS

Figure 5.15: 3D point cloud visualization showing median results from depth maps reconstructed using DVSR and HVSR methods, with and without multiscale filtering, under varying photon count levels (PPP = 1, 4, 16). The first column displays the ground truth, while the last row presents the corresponding noisy inputs.

From left to right, we can observe a clear progression in reconstruction quality as photon count (PPP) increases. The top row displays the ground truth, serving as a reference, while each subsequent row presents the output for DVSR and HVSR before and after applying the multiscale fusion step.
This figure 5.16visually reinforces the quantitative results presented earlier by showcasing median-filtered 3D point cloud reconstructions for both DVSR and HVSR methods across different PPP levels (1, 4, and 16), with and without multiscale filtering.


Ground Truth

PPP = 1

PPP = 4

PPP = 16

GT

DVSR – No MS

DVSR – No MS

DVSR – No MS

DVSR – MS

DVSR – MS

DVSR – MS

HVSR –No MS

HVSR –No MS

HVSR –No MS

HVSR – MS

HVSR – MS

HVSR – MS

Figure 5.16: 3D point cloud visualization showing mean results from histograms reconstructed using DVSR and HVSR methods, with and without multiscale filtering, under varying photon count levels (PPP = 1, 4, 16).

Comments :

- **DVSR** tends to outperform HVSR, producing sharper and more accurate 3D structures, particularly when combined with multiscale pre-processing.
- **Multiscale fusion** significantly improves reconstruction quality, acting as an effective denoising and signal-enhancing step—especially in low-quality settings (e.g., PPP = 1), where inputs are sparse and noisy.
- While all methods tend to converge toward ground truth at higher PPP levels (e.g., PPP = 16), multiscale processing still enhances fine detail consistency.

In summary, the figure provides a strong visual confirmation of the statistical improvements (RMSE, AE) reported earlier, and highlights the robustness of the **multiscale + DVSR** combination, especially when **median fusion is applied to depth maps**.

5.8 Sailency detection results :

In this section, we present UC-Net segmentation results on the DYDTOF dataset, which includes identifiable objects like the cat and detailed scenes suitable for segmentation.

Figure 5.17 illustrates the saliency detection results obtained using the UC-Net algorithm, with bounding boxes highlighting the detected objects. The figure presents four different input scenarios: (1) clean RGB with noisy depth, (2) clean RGB with clean depth (Multiscale + DVSR), (3) noisy RGB with noisy depth, and (4) noisy RGB with clean depth (Multiscale + depth). From left to right, each row shows the input RGB image, the corresponding depth map, and the resulting saliency map.

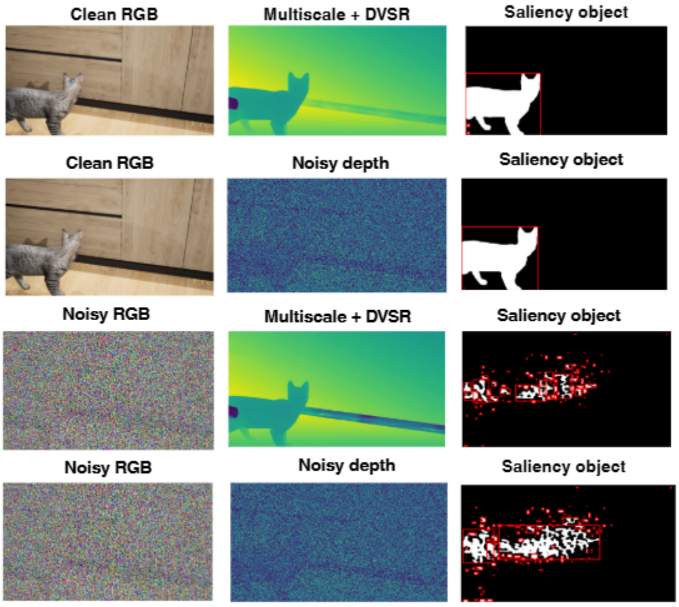


Figure 5.17: Visual results of saliency object detection using UCNet under varying input conditions.

Table 5.5 provides a qualitative comparison of saliency detection outcomes under various combinations of clean and noisy RGB and depth inputs, highlighting the influence of input quality on the model's performance.

Table 5.5: Interpretation of saliency detection under varying input conditions using UC-Net

Case	**RGB Input**	**Depth Input**	**Saliency Detection Output and Interpretation**
1	Clean	Clean	The saliency object is clearly detected, with sharp contours and accurate boundaries. This is the optimal condition, where both RGB and depth data are clean, enabling precise object localization.
2	Clean	Noisy	The object remains detectable, but the saliency mask is slightly degraded. The clean RGB helps compensate for the noisy depth, though spatial coherence is weakened.
3	Noisy	Clean	The model fails to correctly detect the cat, producing a saliency mask that is mostly noisy and does not highlight the object of interest.
4	Noisy	Noisy	The output is heavily degraded, with fragmented detection and poor object localization. With both inputs noisy, the model fails to generate a coherent saliency map.

In summary, the results show that UCNet performs reliably when the RGB input is clean, even in the presence of noisy depth data. However, when the RGB input is degraded—even if the depth map is clean—the model's saliency detection significantly deteriorates. This indicates that UCNet relies more heavily on the RGB modality to produce accurate results.

5.8.1 Computational Cost vs. Performance

In addition to evaluating reconstruction accuracy and robustness, we assessed the computational efficiency of the proposed methods. While deep learning architectures significantly improved the quality of depth estimation, this came at the cost of higher computational demand. For instance, histogram-based methods required longer processing times due to their reliance on high-dimensional data cubes and multi-scale convolution operations.

We observed that:

- Matched Filter methods were faster but less precise, especially under strong noise.
- DVSR and HVSR approaches yielded superior results, particularly with multiscale fusion, but required longer inference times (up to X ms per frame on our setup with GPU XYZ).

- The Plug-and-Play modularity slightly increased overall latency due to sequential module execution, but ensured adaptability and fault tolerance.

5.9 Conclusion

This chapter presented a comparative study of three datasets (Replica, DyDToF, and ARKit), reconstruction methods (Argmax and Matched Filter), multiscale techniques, and saliency detection for depth estimation. The Matched Filter consistently outperformed Argmax, especially under challenging PPP and SBR conditions. Multiscale aggregation using mean or median filters proved effective in reducing noise and improving robustness. Finally, saliency detection enhanced the interpretability of results by highlighting visually or structurally important regions, aiding both analysis and potential downstream tasks.

General Conclusion

General Conclusion

In this work, we explored the intersection of computational imaging, LiDAR technology, and model-based deep learning (MBDL) to enhance the quality and robustness of 3D imaging systems. Beginning with the history and working principles of LiDAR, we highlighted its essential role in modern perception systems, with applications ranging from autonomous navigation to augmented reality. Despite its strengths, LiDAR suffers from limitations such as data sparsity, noise sensitivity, and low resolution under challenging lighting or ambiguous conditions.

To address these issues, we examined state-of-the-art multimodal imaging techniques that fuse LiDAR with complementary data sources, such as RGB images. We introduced MBDL as a new paradigm that combines physical modeling with the adaptability of deep learning. This cross-modal fusion enables us to benefit from the mathematical rigor of physics-based methods and the generalization capabilities of neural networks.

We proposed a new 3D imaging framework that incorporates multiscale histogram-based depth estimation, matched filtering strategies, and a plug-and-play architecture capable of handling various levels of scene complexity and noise. Simulations using noisy photon data allowed us to quantitatively evaluate performance across different datasets using both matched filtering and argmax-based depth recovery methods.

The multiscale approach proved particularly effective. Coarse scales rapidly suppress background noise and provide global structural information, while finer scales enhance edge precision and localization. This hierarchical strategy improves resilience in photon-starved conditions, increases spatial resolution, and significantly reduces root mean square error (RMSE) compared to single-scale methods. Moreover, integrating this approach into both Depth Value Space Reconstruction (DVSR) and Histogram Value Space Reconstruction (HVSR) allowed for efficient depth fusion, particularly in cases with sparse and noisy data.

Experimental results confirmed the effectiveness of our method in managing real-world degradations—especially under low photon counts and poor signal-to-background ratios (SBR)—leading to more accurate depth maps and better scene interpretation. Furthermore, the integration of saliency detection introduced a high-level semantic component, enabling targeted computational focus on visually and contextually important regions.

In summary, this research contributes to the development of robust and intelligent 3D imaging systems by leveraging the synergy between physics-based design and deep learning flexibility. The integration of multiscale matched filtering and argmax estimation into a plug-and-play framework offers a generalizable and real-time-capable solution for multimodal depth perception, particularly in adverse environments. These findings open pathways for future development involving real hardware, live processing, and deployment in robotics, AR/VR, and remote sensing platforms.

Limitations and Future Work

Despite the promising performance of our method, several limitations remain:

- **Dependence on RGB Quality:** The method relies on the fusion of RGB and depth data. In low-light or dark environments, RGB images may be degraded or absent, which negatively impacts the accuracy of depth map reconstruction.
- **Saliency Detection Sensitivity:** The saliency detection module (UCNet) depends strongly on RGB features. When the RGB modality is noisy or missing, salient object detection becomes unreliable, affecting scene understanding.
- **No Testing with Real Hardware:** Our experiments were conducted on three datasets—two synthetic (Replica and DyDToF) and one real (ARKit). However, the method has not yet been tested on real-world data acquired from actual LiDAR and RGB camera systems. Validating it on real sensor data is crucial to assess its robustness in practical scenarios.
- **Limited Modality Robustness:** The current pipeline does not explicitly address scenarios where one modality (e.g., RGB) is missing or corrupted. This could hinder the method's effectiveness in challenging environments such as night-time scenes or adverse weather conditions.

Future Perspectives

- **Generalization to Multi-Sensor Fusion**
  A promising future direction consists in extending the current framework to enable the fusion of data from multiple heterogeneous sensors. In real-world environments, perception systems often depend on diverse sensing modalities such as LiDAR, RGB cameras, thermal imagers, radar, or event-based sensors. Each of these provides complementary information—depth, texture, heat signatures, or motion cues—while differing in terms of resolution, sampling rate, field of view, and noise characteristics. Designing a unified and adaptable fusion strategy would require flexible preprocessing pipelines, alignment of asynchronous data streams, and robust fusion modules capable of handling missing or

degraded inputs. Achieving this generalization would significantly enhance the system's reliability and versatility in complex and dynamic scenarios, particularly in robotics, autonomous navigation, and remote sensing.

- **RGB-D Object Detection Using Depth-Aware Architectures** Another valuable extension of this work involves the development of an object detection framework that leverages both RGB and depth (RGB-D) data. Unlike conventional detectors based solely on RGB imagery, this approach incorporates spatial information to enhance object-background separation, improve localization accuracy, and ensure robustness in visually ambiguous scenes—such as those with poor lighting, clutter, or occlusions. A suitable direction would be to adapt a lightweight and efficient architecture such as YOLO to process RGB-D inputs by modifying the backbone and feature fusion mechanisms to exploit depth-aware cues. The incorporation of depth not only enhances the precision of bounding box placement but also improves semantic understanding of the scene, making this approach particularly useful for applications in robotics, scene understanding, and navigation in structured or unstructured environments.

Bibliography

[1] George Barbastathis, Aydogan Ozcan, and Guohai Situ, *On the Use of Deep Learning for Computational Imaging*,

[2] W. Clem Karl, James E. Fowler, Charles A. Bouman, Müjdat Çetin, Brendt Wohlberg, and Jong Chul Ye, *The Foundations of Computational Imaging: A Signal Processing Perspective*, IEEE Signal Processing Magazine, vol. 40, no. 4, July 2023.

[3] A. Wallace, A. Halimi, and G. S. Buller, "Full Waveform LiDAR for Adverse Weather Conditions," *IEEE Transactions on Vehicular Technology*, in press, 2020.

[4] Kateryna Kuzmenko, Peter Vines, Abderrahim Halimi, Robert J. Collins, Aurora Maccarone, Aongus McCarthy, Zoë M. Greener, Jarosław Kirdoda, Derek C. S. Dumas, Lourdes Ferre Llin, Muhammad M. Mirza, Ross W. Millar, Douglas J. Paul, and Gerald S. Buller, *3D LIDAR imaging using Ge-on-Si single–photon avalanche diode detectors*, Optics Express, vol. 28, no. 2, pp. 1330–1344, January 2020. doi: `10.1364/OE.28.001330`.

[5] Kazuto Nakashima, Yumi Iwashita, and Ryo Kurazume, *Generative Range Imaging for Learning Scene Priors of 3D LiDAR Data*, Kyushu University and Jet Propulsion Laboratory, California Institute of Technology.

[6] Srushti Neoge and Ninad Mehendale, *Review on LiDAR technology*, Noname manuscript, (submitted, under review). [Details like journal name, volume, pages will be added upon acceptance].

[7] J. Tachella, J.-Y. Tourneret, S. McLaughlin, Y. Altmann, N. Mellado, A. McCarthy, R. Tobin, and G. S. Buller, "Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers," in *Nature Communications*, vol. 10, no. 1, 2019.

[8] Ulla Wandinger, *Introduction to Lidar*, Leibniz Institute for Tropospheric Research, Leipzig, Germany.

[9] Zhien Wang and Massimo Menenti, *Challenges and Opportunities in Lidar Remote Sensing*, Frontiers in Remote Sensing, published: 04 March 2021. doi: `10.3389/frsen.2021.641723`.

[10] A. Tontini, S. Mazzucchi, N. Broseghini, R. Passerone, and L. Gasparini, "Histogram-less LiDAR through SPAD response linearization," *IEEE Sensors Journal*, Feb. 2024.

[11] K. Kuzmenko, P. Vines, A. Halimi, R. J. Collins, A. Maccarone, A. McCarthy, Z. M. Greener, J. Kirdoda, D. C. S. Dumas, L. Ferre Llin, M. M. Mirza, R. W. Millar, D. J. Paul, and G. S. Buller, "3D LiDAR imaging using Ge-on-Si single–photon avalanche diode detectors," *Research Article*, Heriot-Watt University and University of Glasgow, Available from corresponding author: `G.S.Buller@hw.ac.uk`.

[12] J. Koo, A. Halimi, and S. McLaughlin, “A Bayesian based deep unrolling algorithm for single-photon LiDAR systems,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 762–777, June 2022.

[13] Z. Li, Y. Han, L. Wu, Z. Zang, M. Dai, S. Y. Set, S. Yamashita, Q. Li, and H. Y. Fu, “Towards an ultrafast 3D imaging scanning LiDAR system: a review,” *Photonics Research*, vol. 12, no. 8, pp. 1709–1726, Aug. 2024. Available at: https://opg.optica.org/prj/fulltext.cfm?uri=prj-12-8-1709&id=539262,

[14] D. Yao, G. Mora-Martín, I. Gyongy, S. Scholes, J. Leach, S. McLaughlin, and Y. Altmann, “Bayesian neuromorphic imaging for single-photon LiDAR,” *Research Article*, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, UK,

[15] https://www.androidpolice.com/what-is-lidar-and-what-does-it-do/

[16] Pratiksha Pendse, *An Overview on LiDAR for Autonomous Vehicles*, Department of Computer Science, Technische Universität Chemnitz, Germany, April 2024. doi: 10.5281/zenodo.10992391. Available at: https://www.researchgate.net/publication/379900256.

[17] Santiago Royo and Maria Ballesta-Garcia, *An Overview of Lidar Imaging Systems for Autonomous Vehicles*, Sciences, Centre for Sensor, Instrumentation and Systems Development (CD6), Universitat Politècnica de Catalunya, published: 30 September 2019.

[18] Y. Li, H. Guo, and H. Sheng, “Self-supervised single-view 3D point cloud reconstruction through GAN inversion,” *The Journal of Supercomputing*, vol. 80, pp. 21365–21393, 2024.

[19] L. Yu, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, “PU-Net: Point Cloud Upsampling Network,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2790–2799. Available at: https://ieeexplore.ieee.org/document/8578393,

[20] G. Qian, A. Abualshour, G. Li, A. Thabet, and B. Ghanem, “PU-GCN: Point Cloud Upsampling using Graph Convolutional Networks, King Abdullah University of Science and Technology (KAUST), Available at: https://arxiv.org/abs/2103.07533,

[21] Y. Li, S. Xie, Z. Wan, H. Lv, H. Song, and Z. Lv, “Graph-powered learning methods in the Internet of Things: A survey,” *Machine Learning with Applications*, vol. 11, 100441, 2023. Available at: https://www.sciencedirect.com/science/article/pii/S2666827023000201,

[22] W. Yookwan, K. Chinnasarn, C. So-In, and P. Horkaew, “Multimodal Fusion of Deeply Inferred Point Clouds for 3D Scene Reconstruction Using Cross-Entropy ICP,” *IEEE Access*, vol. 10, pp. 82166–82180, doi: 10.1109/ACCESS.2022.3192869, Received: Jun. 29, 2022; Accepted: Jul. 17, 2022; Published: Jul. 21, 2022; Current version: Jul. 27, 2022.

[23] J. C. Fernandez, A. Singhania, J. Caceres, K. C. Slatton, M. Starek, and R. Kumar, “An Overview of Lidar Point Cloud Processing Software,” GEM Center Report No. Rep_2007-12-001, Geosensing Engineering and Mapping (GEM), Civil and Coastal Engineering Department, University of Florida, Dec. 20, 2007. Available at: https://www.essie.ufl.edu/,

[24] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, “UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8582–8591, June 2020. Available at: https://arxiv.org/abs/2004.05763,

[25] S. Kalamkar, A. Geetha Mary, “Multimodal Image Fusion: A Systematic Review,” in *Visual Informatics*, vol. 7, no. 4, pp. 1–20, Dec. 2023. Available at: https://www.sciencedirect.com/science/article/pii/S2772662223001674,

[26] R. R. Nair, T. Singh, R. Sankar, and K. Gunndu, “Multi-modal medical image fusion using LMF-GAN – A maximum parameter infusion technique,” *Journal of Intelligent & Fuzzy Systems*2021.

[27] Z. Sun, W. Ye, J. Xiong, G. Choe, J. Wang, S. Su, R. Ranjan, “Consistent Direct Time-of-Flight Video Depth Super-Resolution,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Sun_Consistent_Direct_Time-of-Flight_Video_Depth_Super-Resolution_CVPR_2023_paper.html,

[28] J. Park, K. Joo, Z. Hu, C.-K. Liu, and I. S. Kweon, “Non-local Spatial Propagation Network for Depth Completion,” *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 120–136, 2020. Available at: https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123580120.pdf,

[29] Y. Wang, S. Wu, H. Huang, D. Cohen-Or, and O. Sorkine-Hornung, “Patch-based Progressive 3D Point Set Upsampling,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Available at: https://arxiv.org/abs/1811.11286,

[30] J. Zhang, H. Chen, Z. Liu, J. Ma, and X. Mei, “Transformer Based Conditional GAN for Multimodal Image Fusion,” *IEEE Transactions on Multimedia*, vol. 25, pp. 8988–9001, 2023, doi: 10.1109/TMM.2023.3243659.

[31] A. Halimi, A. Maccarone, R. Lamb, G. Buller, and S. McLaughlin, “Robust and Guided Bayesian Reconstruction of Single-Photon 3D LiDAR Data: Application to Multispectral and Underwater Imaging,” *IEEE Transactions on Computational Imaging*, 2021.

[32] L. Li, Z. Zhu, and C. Wang, “Multiscale Entropy-Based Surface Complexity Analysis for Land Cover Image Semantic Segmentation,”

[33] A. Ruget, L. Wilson, J. Leach, R. Tobin, A. McCarthy, G. S. Buller, S. McLaughlin, and A. Halimi, “A Plug-and-Play Algorithm for 3D Video Super-Resolution of Single-Photon LiDAR Data,” *arXiv preprint*, 2024.