

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова
праця на правах рукопису

САВІН ВОЛОДИМИР ВАДИМОВИЧ

УДК 004.8, 004.93

ДИСЕРТАЦІЯ
МАТЕМАТИЧНІ МОДЕЛІ ТА МЕТОДИ 3D РЕКОНСТРУКЦІЇ В
ДОПОВНЕНІЙ РЕАЛЬНОСТІ

Спеціальність 113 «Прикладна математика»

Галузь знань 11 «Математика та статистика»

Подається на здобуття наукового ступеня доктора філософії.

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ Савін В. В.

Науковий керівник Куссуль Наталія Миколаївна, доктор технічних наук, професор.

Київ – 2026

АНОТАЦІЯ

Savin B.B. Математичні моделі та методи 3D реконструкції в доповненій реальності. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня доктора філософії за спеціальністю 113 «Прикладна математика» (галузь знань 11 «Математика та Статистика»). – Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, 2026.

В останні десятиліття спостерігається надзвичайно швидкий розвиток технологій доповненої реальності (ДР), які відкривають новий етап взаємодії людини з інформацією та оточуючим середовищем. Сценарії використання цих технологій стрімко еволюціонують, переходячи з мобільних на самостійні-носимі пристрої. Продукти, які дозволяють споживати сценарії ДР, виходять зі стадії обмежених та дорогих платформ розробки і набувають масовості.

Сучасні тренди у галузі доповненої реальності визначаються не лише зростанням її популярності, але й постійним удосконаленням технологічного арсеналу. Все більше компаній і дослідницьких груп зосереджують свою увагу на розробці інноваційних методів та рішень спрямованих на підвищення реалістичності взаємодії об'єктів ДР з реальним світом та користувача з об'єктами ДР.

Одні з ключових технологій, необхідних для досягнення максимальної імерсії взаємодії та реалізації потенціалу ДР є реконструкція карт глибини та 3D реконструкція оточуючого середовища і його об'єктів. Мета 3D реконструкції полягає у відтворенні тривимірної моделі оточення на основі доступних даних: дво- та тривимірні зображення, внутрішні та зовнішні параметри камери, дані з IMU сенсорів, або систем позиціонування і т.д. 3D реконструкція є своєрідним каркасом для доповненої реальності, а її основною складовою є отримання (за рахунок апаратних засобів), чи відтворення (за рахунок алгоритмів на основі даних з одної чи декількох камер) карт глибини. Якість моделі реконструйованої сцени суттєво впливає на реалістичність сприйняття сценаріїв доповненої реальності користувачем. Справжній виклик полягає в тому, як забезпечити

точність, достовірність та ефективність 3D реконструкції середовища при виконанні алгоритмів на кінцевих пристроях користувача та апаратних прискорювачах з обмеженою розрядністю (DSP/NPU), особливо, в умовах наявності на сцені динамічного освітлення та складних поверхонь (відбивні чи напівпрозорі).

Метою даного дослідження є підвищення якості, стійкості та енергоефективності методів 3D реконструкції середовища, що виконуються на кінцевому пристрої користувача, для забезпечення реалістичної та надійної взаємодії користувача з віртуальними об'єктами в системах доповненої реальності.

Якість існуючих класичних підходів реконструкції карт глибини чи 3D реконструкції сцени недостатня для реалізації якісних та імерсивних сценаріїв доповненої реальності. Такі алгоритми часто помиляються на складних поверхнях, таких як: однорідні, прозорі/напівпрозорі, відбивні, поверхні з періодичною текстурою; та на складних сценах, що містять: тонкі чи дрібні структури, динамічне освітлення, динамічні об'єкти. Сучасні методи реконструкції карт глибини чи 3D реконструкції сцени вирішують, або мінімізують вплив вищезазначених складних сцен та поверхонь. Але їх обчислювальна складність робить їх непридатними, або не ефективними для використання на кінцевих пристроях користувача з обмеженими обчислювальними можливостями та з батареями невисокої ємності.

Тому актуальною є розробка нових методів та підходів, що дозволять врахувати особливості складних сцен при їх реконструкції на кінцевих пристроях користувача та на апаратних прискорювачах з обмеженою розрядністю.

Перший розділ дисертаційного дослідження систематизує популярні існуючі роботи та підходи, що присвячені задачам реконструкції карт глибини та 3D реконструкції сцени. Демонструються результати аналізу даних методів, визначено їх основні недоліки та сформульовано завдання дисертаційного дослідження.

У *другому* розділі досліджується проблема реконструкції середовища в умовах динамічного освітлення сцени. Запропоновано модифікацію Neural Radiance Fields (NeRF) моделі до умов динамічного освітлення шляхом модифікації функцій втрат та введення часової змінної. Проведені експерименти на відкритих наборах даних, які демонструють ознаки наявності динамічного освітлення, підтверджують ефективність запропонованого підходу. Також, проведено аналіз та запропоновано використання описаного методу для розширення (аугментації) наявних наборів даних шляхом синтезу зображень новітніх видів сцени з метою подальшого навчання легших моделей орієнтованих на ефективне виконання на кінцевому пристрої користувача.

У *третьому розділі* дисертаційної роботи розглядається проблема реконструкції карт глибини при наявних на сцені відбивних чи напівпрозорих поверхонь. Запропоновано метод, який дозволяє отримати та окремо зберігти значення глибини до самої складної поверхні та до відбитого чи перекритого об'єкту. Описаний метод запатентовано як винахід в глобальній базі WIPO PATENTSCOPE та присутній в патентних базах багатьох країн, включаючи Республіка Корея, США та ін.

Четвертий розділ адресує проблему ефективної реконструкції карт глибини високої точності та з широким діапазоном глибини на апаратних прискорювачах з обмеженою розрядністю кінцевих пристроїв користувача. Запропоновано метод, особливістю якого є модифікація архітектури моделі реконструкції карт глибини шляхом додавання виходів для окремого передбачення компонент двовимірної кривої Гільберта. Виходи моделі піддаються простій постобробці для отримання карт глибини широкого діапазону. Представлені результати експериментів на відкритих наборах даних демонструють:

- суттєве зменшення похибки квантування моделі, виконання якої планується на апаратних прискорювачах з обмеженою розрядністю;
- покращення якості реконструкції карт глибини, особливо при наявності дрібних структур;

- збільшення розрядності результуючих значень карт глибини, що призводить до розширення робочого діапазону та точності значень глибини;
- зменшення часу виконання та енергоспоживання результуючої квантованої моделі.

Демонструються результати аналізу обмежень запропонованого підходу та можливості для його подальшого розвитку.

Наукова новизна отриманих результатів:

1. Удосконалено метод Neural Radiance Fields (NeRF) до умов динамічного освітлення шляхом модифікації функцій втрат та введення часової змінної, що дозволило покращити якість реконструкції сцен з динамічним освітленням та розширювати існуючі набори даних для складних сцен.
2. Вперше розроблено та запатентовано метод реконструкції глибини, який враховує напівпрозорі та відбивні поверхні, зберігає окремо значення глибини до самої площини та до відбитого/перекритого об'єкту, що дозволяє збільшити точність реконструкції складних сцен, які містять не дифузні поверхні.
3. Вперше сформульовано та розроблено метод для прогнозування карт глибини з представленням виходу моделі у вигляді компонент двовимірної кривої Гільберта, що дозволило для квантованих моделей розширити діапазон глибини, підвищити її точність та покращити енергоефективність реконструкції сцени на апаратних прискорювачах з обмеженою розрядністю (DSP/NPU) кінцевих пристроїв користувача.

Практичне значення отриманих результатів та їх застосування.

Результати присутні в даному дослідженні були використані в рамках комерційних та науково-дослідницьких проектів ТОВ «Самсунг РнД Інститут Україна». Запропоновані методи знайшли застосування у галузі візуального

інтелекту при вирішенні завдань 3D реконструкції сцени для доповненої реальності та створення/редагування просторового контенту.

1. Метод призначений для реконструкції карт глибини з урахуванням напівпрозорих та відбивних поверхонь, що орієнтований на підвищення якості відтворення складних об'єктів, зокрема при наявності динамічного освітлення захищено патентом (US20240144503A1). Отриманий патент розширює патентне портфоліо компанії у відповідному технологічному домені.
2. Метод прогнозування високоточних карт глибини з широким діапазоном на пристроях з обмеженою розрядністю, що ґрунтується на використанні двовимірних кривих Гільберта, дозволяє: зменшити похибку квантування моделі у 4,6 рази, що підвищує якість реконструкції карт глибини на DSP-пристроях; розширити діапазон відстаней за рахунок збільшення ефективної розрядності з 8-біт до 10-біт; зменшити час виконання та енергоспоживання квантованої моделі у 1,5 раза порівняно з оригінальною моделлю за умови збереження або покращення якості прогнозування карт глибини.
3. Розглянуті та запропоновані методи 3D реконструкції сцени на кінцевих пристроях користувача з обмеженими обчислювальними ресурсами було використано при розробці комерційних проектів, що спрямовані на сценарії створення/редагування просторового контенту для флагманської моделі смартфона Samsung Galaxy S25.

Також, представлення виходу моделі на основі кривої Гільберта, що запропоновано у методі прогнозування карт глибини високої точності з широким діапазоном на пристроях з обмеженою розрядністю може бути застосовано для покращення якості та ефективності і інших задач комп'ютерного зору на основі методів машинного навчання, що виконуються на апаратних прискорювачах з низькою арифметикою.

Ключові слова: машинне навчання, глибинне навчання, нейронна мережа, комп'ютерний зір, карта глибини, 3D реконструкція, Neural Radiance Fields,

квантування, криві Гільберта, виконання моделі на кінцевому пристрої, генерація навчальних даних, аугментація, функція втрат, DispNet, Dense Prediction Transformer.

ABSTRACT

Savin V.V. Mathematical models and methods of 3D reconstruction in Augmented Reality. – A qualification research work, manuscript.

PhD thesis in specialty 113 “Applied Mathematics” (field of knowledge 11 “Mathematics and Statistics”). – National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute,” Kyiv, 2026.

In recent decades, augmented reality (AR) technologies have advanced at an exceptional pace, opening a new phase of human interaction with information and the surrounding environment. Usage scenarios are rapidly evolving from mobile to standalone wearable devices. Products enabling AR experiences are moving from the stage of limited and expensive developer platforms towards mass adoption.

Current trends in AR are defined not only by growing popularity but also by continuous enhancement of the technological toolkit. An increasing number of companies and research groups focus on developing innovative methods and solutions that raise the realism of interactions between AR objects and the real world, as well as between users and AR content.

Among the key technologies required to achieve highly immersive interaction and to enable the potential of AR are depth-map estimation and 3D reconstruction of the environment and its objects. The goal of 3D reconstruction is to recover a three-dimensional model of a scene from available data: 2D/3D images, camera intrinsics and extrinsics, IMU sensor signals, positioning systems, etc. 3D reconstruction forms the structural backbone of AR. Its core component is the acquisition (via hardware) or estimation (via algorithms operating on data from one or multiple cameras) of depth maps. The quality of the reconstructed scene model critically affects how the user perceives the realism of AR scenarios. A main challenge is ensuring the accuracy, fidelity, and efficiency of 3D reconstruction on end-user devices and low-precision hardware accelerators (DSP/NPU) in the presence of dynamic lighting and challenging surface types (reflective or semi-transparent).

The aim of this work is to improve the quality, stability and energy efficiency of on-device 3D environment reconstruction methods, in order to enable realistic and reliable user interaction with virtual objects in augmented reality systems.

The quality of existing classical approaches of depth-map estimation and 3D scene reconstruction is insufficient for high-quality, immersive AR experiences. Such algorithms frequently fail on challenging surfaces, including homogeneous, transparent/semi-transparent, reflective, and periodically textured materials, as well as in complex scenes containing thin or fine structures, dynamic illumination, and dynamic objects. Modern methods address or mitigate many of these challenges, but their computational cost makes them impractical or inefficient on resource-constrained end devices with limited battery capacity.

Therefore, the development of new methods and approaches that enables efficient on-device scene reconstruction at low-precision hardware accelerators, remains a relevant research task.

Chapter 1 systematizes prominent existing works and approaches devoted to depth-map estimation and 3D scene reconstruction. It presents an analysis of these methods, identifies their main limitations, and formulates the research objectives of the PhD thesis.

Chapter 2 studies the problem of environment reconstruction considering dynamic illumination conditions. A modification of Neural Radiance Fields (NeRF) is proposed to enable scene reconstruction under dynamic illumination conditions by adjusting the loss functions and introducing a temporal variable. Experiments on open datasets exhibiting signs of dynamic illumination confirm the effectiveness of the proposed approach. The chapter also analyzes and proposes the use of this method for augmenting existing datasets by synthesizing novel views of a scene for further lighter models training focused at efficient on-device inference.

Chapter 3 addresses depth-map reconstruction in scenes containing reflective or semi-transparent surfaces. A method is proposed that enables obtaining and separately storing the depth value to the complex surface itself and to the reflected or occluded object. The described method has been patented as an invention in the global WIPO

PATENTSCOPE database and is also registered in the patent databases of many countries, including the Republic of Korea, the United States, and others.

Chapter 4 tackles efficient reconstruction of high-precision, wide-range depth maps on low-precision hardware accelerators of end devices. The key feature of proposed method is a modification of the depth-estimation model architecture by adding outputs that separately predict the components of a two-dimensional Hilbert curve. The model outputs are then post-processed with a simple procedure to obtain wide-range depth maps. Experiments on open datasets demonstrate:

- a significant reduction of the model quantization error, enabling its accurate inference on low-precision accelerators;
- improved depth-map reconstruction quality, especially for fine structures;
- increased effective bit-depth of the resulting depth values, expanding the operating range and depth accuracy;
- reduced inference time and energy consumption of the resulting quantized model.

The chapter also presents an analysis of the proposed approach's limitations and outlines opportunities for further development.

Scientific novelty of the obtained results:

1. The Neural Radiance Fields (NeRF) method was improved for dynamic lighting conditions by modifying the loss functions and introducing a temporal variable, which made it possible to enhance reconstruction quality for scenes with dynamic illumination and to expand existing datasets for complex scenes.
2. For the first time, a depth-reconstruction method was developed and patented that accounts for semi-transparent and reflective surfaces and separately stores the depth value to the surface plane itself and to the reflected/occluded object, thereby increasing reconstruction accuracy for complex scenes containing non-diffuse surfaces.
3. For the first time, a method was formulated and developed for depth-map prediction in which the model output is represented as components of a

two-dimensional Hilbert curve. This enabled quantized models to extend the depth range, improve depth accuracy, and increase the energy efficiency of scene reconstruction on low-precision hardware accelerators (DSP/NPU) of end-user devices.

Practical significance of the obtained results and their application.

The results presented in this study were used within commercial and R&D projects of Samsung R&D Institute Ukraine LLC. The proposed methods have been applied in the visual intelligence domain for solving problems of 3D scene reconstruction for augmented reality and for spatial content creation/editing.

1. A method for depth-map reconstruction that accounts for semi-transparent and reflective surfaces and is aimed at improving the reconstruction quality of complex objects, including under dynamic lighting, is protected by a patent (US20240144503A1). The granted patent expands the company's patent portfolio in the corresponding technological domain.
2. A method for predicting high-precision, wide-range depth maps on low-precision devices, based on the use of two-dimensional Hilbert curves, enables: a $4.6\times$ reduction in model quantization error, which improves depth-map reconstruction quality on DSP devices; an extension of the distance range by increasing the effective bit-depth from 8-bit to 10-bit; and a $1.5\times$ reduction in inference time and energy consumption of the quantized model compared to the original model, while maintaining or improving depth-prediction quality.
3. The considered and proposed methods for on-device 3D scene reconstruction under limited computational resources were used in the development of commercial projects targeting spatial content creation/editing scenarios for the flagship smartphone model Samsung Galaxy S25.

Furthermore, representing the model output based on a Hilbert curve, as proposed in the method for high-precision, wide-range depth-map prediction on low-precision devices, can also be applied to improve the quality and efficiency of other

computer-vision tasks based on machine-learning methods executed on low-precision hardware accelerators.

Keywords: machine learning, deep learning, neural network, computer vision, depth map, 3D reconstruction, Neural Radiance Fields, quantization, Hilbert curves, on-device inference, training-data generation, augmentation, loss function, DispNet, Dense Prediction Transformer.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. Savin V., Kolodiazhna O., Adapting Neural Radiance Fields (NeRF) to the 3D Scene Reconstruction Problem Under Dynamic Illumination Conditions, *Cybernetics and Systems Analysis*, 2023, vol. 59, pp. 910 – 918, ISSN: 1060-0396, DOI: 10.1007/s10559-023-00626-7. [Scopus, WoS, Q3]

Внесок співавторів:

- Savin V.: сформульовано проблему та мотивацію роботи; проведено огляд літератури в області використання Neural Radiance Fields для задач 3D реконструкції та синтезу наборів даних; підготовлено науковий контекст задачі; підготовлено дані для експериментів (відбір сцен/кадрів, попередня обробка, формування навчальних та валідаційних підмножин); здійснено підбір і валідацію гіперпараметрів навчання; запропоновано і проведено експерименти та порівняльний аналіз отриманих результатів, що підтвердило перевагу запропонованого підходу над оригінальною моделлю для задач 3D реконструкції сцени в умовах динамічного освітлення; написання тексту статті; проведено роботу із зовнішніми рецензентами.
- Kolodiazhna O.: сформульовано проблему та мотивацію роботи; проведено постановку задачі; обрано підхід та методологію; розроблено метод, та відповідні алгоритми; розроблено загальну стратегію експериментів; обрано та обґрунтовано метрики; реалізовано метрики та скрипти оцінювання; проведено внутрішнє рецензування.

2. Uss M., Iermolenko R., Kolodiazhna O., Savin V., Method and device for generating depth map, Patent: US20240144503A1, URL: <https://patents.google.com/patent/US20240144503A1>.

Внесок співавторів:

- Uss M.: розширено оригінальну ідею та запропоновано їй технічну деталізацію для патентної заявки; сформульовано проблему та мотивацію роботи; проведено постановку задачі; обрано підхід та методологію, розроблено загальну стратегію експериментів; розроблено архітектуру та метод; оформлено відповідні розділи патентної заявки.
- Iermolenko R., Kolodiazhna O.: реалізовано прототип; обрано набори даних та сценарії тестування; обрано та обґрунтовано метрики; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.
- Savin V.: проведено аналіз існуючих робіт та винаходів в області реконструкції глибини враховуючи не дифузні поверхні; запропоновано ідею методу реконструкції глибини, який враховує напівпрозорі та відбивні поверхні; запропоновано та проведено ряд експериментів для підтвердження доцільності методу реконструкції глибини, який враховує напівпрозорі та відбивні поверхні, зберігає окремо значення глибини до самої площини та до відбитого/перекритого об'єкту, що дозволяє збільшити точність реконструкції складних сцен; оформлено відповідні розділи патентної заявки; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.

3. Савін В.В., Аналіз методів 3D реконструкції середовища для доповненої реальності, *Методи комп'ютерного зору і глибинних нейронних мереж для еколого-економічного аналізу*: монографія / за ред. Н.М. Куссуль, А.Ю. Шелестова – Київ: Наукова думка, 2024. – 448с. С. 49 – 81, ISBN 978-966-00-1940-9.

4. Kolodiazhna O., Savin V., Uss M., Kussul N., 3D Scene Reconstruction with Neural Radiance Fields (NeRF) considering dynamic illumination conditions,

Proceedings of International Conference on Applied Innovation in IT 2023, 2023, Volume 11, Issue 1, pp. 233-238. ISSN: 2199-8876, DOI: 10.25673/101943. [Scopus]

Внесок співавторів:

- Kolodiazhna O.: проведено постановку задачі; обрано підхід та методологію, розроблено загальну стратегію експериментів; розроблено метод та відповідні алгоритми; обрано та обґрунтовано метрики; реалізовано метрики та скрипти оцінювання.
- Savin V.: сформульовано проблему та мотивацію роботи; проведено огляд літератури в області використання Neural Radiance Fields для задач 3D реконструкції та синтезу наборів даних; підготовлено науковий контекст задачі; підготовлено дані для експериментів (відбір сцен/кадрів, попередня обробка, формування навчальних та валідаційних підмножин); здійснено відбір і валідацію гіперпараметрів навчання; запропоновано і проведено експерименти та порівняльний аналіз отриманих результатів, що підтвердило перевагу запропонованого підходу над оригінальною моделлю для задач 3D реконструкції сцени в умовах динамічного освітлення.
- Uss M., Kussul N.: проведено внутрішнє рецензування.

5. Uss M., Iermolenko R., Shashko O., Kolodiazhna O., Safonov I., Savin V., Yeo Y, Ji S., Jeong J., Predicting High-precision Depth on Low-Precision Devices Using 2D Hilbert Curves, *Proceedings of the 42nd International Conference on Machine Learning*, PMLR, 2025, vol. 267, pp. 60635 – 60656, ISSN: 2640-3498. [Scopus]

Внесок співавторів:

- Uss M.: сформульовано основну ідею; проведено постановку задачі та мотивацію роботи; проведено планування експериментів та аналіз результатів; розроблено функцію втрат; підготовлено чорновий варіанту статті.
- Iermolenko R.: проведено роботу з набором даних; проведено квантування моделі; рішення розгорнуто на пристрої.

- Kolodiazhna O.: проведено ряд експериментів з DispNet моделлю; проведено ряд експериментів з HPE рішенням; проведено ряд експериментів на наборі даних KITTI; підготовлено чорновий варіант статті.
- Shashko O.: проведено ряд експериментів з DPT моделлю; проведено ряд експериментів з HPE рішенням; підготовлено чорновий варіант статті.
- Savin V.: проведено аналіз робіт попередніх дослідників орієнтованих на підвищення точності та енергоефективності реконструкції сцени; удосконалено оригінальну ідею та прийнято участь в розробці методу її технічної реалізації; проведено аналіз та запропоновано оптимальну параметричну криву та її порядок, що дозволило реалізувати метод прогнозування карт глибини з представленням виходу моделі у вигляді компонент двовимірної кривої Гільберта; запропоновано та реалізовано зворотне перетворення результатів виходу моделі та алгоритм постобробки, що сприяло розширенню діапазону глибини для квантованої моделі при її виконанні на прискорювачах з обмеженою розрядністю (DSP/NPU), підвищенню її точності та енергоефективності; обрано метрики оцінки якості запропонованого підходу; проведено ряд експериментів, що підтверджують переваги запропонованого підходу; підготовлено чорновий варіант статті.
- Safonov I.: проведено аналіз альтернатив кривим, що заповнюють простір; підготовлено додаткові матеріали; підготовлено чорновий варіант статті.
- Jeong J., Ji S., Yeo Y: проведено внутрішнє рецензування.

6. Савін В. В., Блохіна І. О., Використання технологій доповненої та віртуальної реальності в умовах дистанційного навчання, Наука та освіта в дослідженнях молодих учених: матеріали IV Міжнар. наук.-практ. конф. для студ., аспірантів, докторантів, молодих учених, Харків, 18 трав. 2023 р. / Харків.

нац. пед. ун-т ім. Г. С. Сковороди., URL: <https://dspace.hnpu.edu.ua/items/25b98239-a99a-476b-b170-029b6b10f161>.

Внесок співавторів:

- Савін В. В.: сформульовано проблему та мотивацію роботи; проведено постановку задачі; підготовлено текст статті.
- Блохіна І. О.: рецензування роботи.

7. Marchenko A., Savin V., Tymchyshyn V., Gesture sensing method and electronic device supporting same, Patent: US20180329501A1, URL: <https://patents.google.com/patent/US20180329501A1>.

Внесок співавторів:

- Marchenko A.: запропоновано оригінальну ідею; проведено постановку задачі; обрано підхід та методологію; розроблено метод; реалізовано частини прототипу; оформлено відповідні розділи патентної заявки.
- Savin V.: сформульовано проблему та мотивацію роботи; доопрацьовано та вдосконалено оригінальну ідею; проведено аналіз існуючих робіт та винаходів; розроблено загальну стратегію експериментів; реалізовано частини прототипу; оформлено відповідні розділи патентної заявки; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.
- Tymchyshyn V.: обрано набори даних та сценарії тестування; обрано та обґрунтовано метрики; реалізовано частини прототипу; проведено ряд експериментів для підтвердження доцільності запропонованого методу.

8. Sydorenko D., Alkhimova S., Savin V., Shcherbina A., Kim G., Bondarets I., Electronic device and method for identifying relevant device in augmented reality mode of electronic device, Patent: US20220070431A1, URL: <https://patents.google.com/patent/US20220070431A1>.

Внесок співавторів:

- Sydorenko D., Alkhimova S., Bondarets I.: доопрацьовано та вдосконалено оригінальну ідею; обрано підхід та методологію; розроблено метод; реалізовано прототип; розроблено загальну стратегію експериментів; обрано набори даних та сценарії тестування; обрано та обґрунтовано метрики; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено патентну заявку; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.
- Savin V.: запропоновано оригінальну ідею; проведено постановку задачі; сформульовано проблему та мотивацію роботи.
- Shcherbina A.: проведено аналіз існуючих робіт та винаходів.
- Kim G.: обговорено доцільність та вплив запропонованого методу.

9. Shcherbina A., Bondarets I., Trunov O., Olshevskiy V., Savin V., Method of adaptive 6DoF hand parameters estimation for precise interaction in AR, Patent: US20230004214A1, URL: <https://patents.google.com/patent/US20230004214A1>.

Внесок співавторів:

- Shcherbina A.: запропоновано оригінальну ідею; проведено постановку задачі; обрано підхід та методологію; розроблено метод; розроблено загальну стратегію експериментів; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.
- Bondarets I., Trunov O., Olshevskiy V.: доопрацьовано та вдосконалено оригінальну ідею; проведено аналіз існуючих робіт та винаходів; розроблено частини методу; реалізовано прототип; обрано та обґрунтовано метрики; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.
- Savin V.: сформульовано проблему та мотивацію роботи; розроблено частини методу; обрано набори даних та сценарії тестування;

оформлено відповідні розділи патентної заявки; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.

10. Iermolenko R., Sukhariev A., Morozov K., Vdovychenko I., Savin V., Vavdiiuk D., Klimenkov O., Sapozhnik O., Electronic apparatus and controlling method thereof, Patent: US20230254568A1, URL: <https://patents.google.com/patent/US20230254568A1>.

Внесок співавторів:

- Iermolenko R., Sukhariev A.: доопрацьовано та вдосконалено оригінальну ідею; обрано підхід та методологію; розроблено метод; розроблено загальну стратегію експериментів; реалізовано частини прототипу; оформлено відповідні розділи патентної заявки.
- Morozov K.: запропоновано оригінальну ідею; проведено постановку задачі; розроблено метод; реалізовано частини прототипу; оформлено відповідні розділи патентної заявки.
- Vdovychenko I.: проведено аналіз існуючих робіт та винаходів; реалізовано частини прототипу; проведено ряд експериментів для підтвердження доцільності запропонованого методу; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.
- Savin V.: сформульовано проблему та мотивацію роботи; обрано та обґрунтовано метрики; обрано набори даних та сценарії тестування; проведено ряд експериментів для підтвердження доцільності запропонованого методу;
- Vavdiiuk D., Klimenkov O., Sapozhnik O.: доопрацьовано та вдосконалено оригінальну ідею; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.

11. Vdovychenko I., Sapozhnik O., Dykyi V., Savin V., Vitiuk A., Tuzhykov A., Electronic device and method for providing augmented reality environment

including adaptive multi-camera, Patent: US20240104867A1, URL: <https://patents.google.com/patent/US20240104867A1>.

Внесок співавторів:

- Vdovychenko I., Sapozhnik O.: доопрацьовано та вдосконалено оригінальну ідею; розроблено метод; розроблено загальну стратегію експериментів; реалізовано частини прототипу; оформлено відповідні розділи патентної заявки.
- Dykui V.: проведено аналіз існуючих робіт та винаходів; обрано та обґрунтовано метрики; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.
- Savin V.: запропоновано оригінальну ідею; проведено постановку задачі; сформульовано проблему та мотивацію роботи; обрано підхід та методологію; розроблено метод; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту; оформлено відповідні розділи патентної заявки.
- Vitiuk A., Tuzhykov A.: обрано набори даних та сценарії тестування; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.

12. Omelchenko A., Vdovychenko I., Morozov K., Androsov V., Savin V., Electronic device for controlling audio device on basis of image context, and method for operating same, Patent: US20250193598A1, URL: <https://patents.google.com/patent/US20250193598A1>.

Внесок співавторів:

- Omelchenko A.: запропоновано оригінальну ідею; проведено постановку задачі; сформульовано проблему та мотивацію роботи; обрано підхід та методологію; розроблено метод; розроблено загальну стратегію експериментів; реалізовано частини прототипу;

обрано та обґрунтовано метрики; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.

- Vdovychenko I., Morozov K., Androsov V.: доопрацьовано та вдосконалено оригінальну ідею; проведено аналіз існуючих робіт та винаходів; реалізовано частини прототипу; обрано набори даних та сценарії тестування; проведено ряд експериментів для підтвердження доцільності запропонованого методу; оформлено відповідні розділи патентної заявки.
- Savin V.: доопрацьовано та вдосконалено оригінальну ідею; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ	26
ВСТУП.....	28
РОЗДІЛ 1: АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА ПІДХОДІВ 3D РЕКОНСТРУКЦІЇ СЦЕНИ	35
1.1. Методи реконструкції карти глибини по 2D даним.....	41
1.1.1. Лінійна перспектива	42
1.1.2. Атмосферне розсіювання	43
1.1.3. Розуміння глибини по затіненню	43
1.1.4. Бінокулярна диспаратність.....	44
1.1.5. Паралакс руху	45
1.1.6. Розмиття зображення.....	46
1.2. Класичні методи 3D реконструкції середовища	47
1.2.1. Відтворення форми за силуетом	48
1.2.2. Структура з руху	49
1.2.3. Багатовидове стерео співставлення	52
1.2.4. Реконструкція поверхні	54
1.2.5. Підхід, який поєднує SfM, MVS та Реконструкцію поверхні	55
1.3. Сучасні методи реконструкції карт глибини та 3D реконструкції середовища	56
1.3.1. HighRes-MVSNet	57
1.3.2. 3D-FHNet.....	59
1.3.3. ATLAS	61
1.3.4. SimpleRecon	62
1.3.5. Marigold.....	64

	23
1.4. Методи реконструкції та рендерингу	66
1.4.1. Neural Radiance Fields (NeRF)	66
1.4.2. 3D Gaussian Splatting (3DGS)	68
1.4.3. NeRF похідні методи	72
1.5. Сучасні підходи реконструкції карт глибини та 3D реконструкції середовища орієнтовані на кінцеві пристрої	74
1.5.1. DispNet	74
1.5.2. LightStereo	76
1.5.3. AnyNet	77
1.5.4. Метод зниження вимог по пам'яті для доповнення карт глибини квантованою мережею	80
1.5.5. MobileViTv2	82
1.6. Порівняльний аналіз методів та постановка задачі дослідження	84
1.7. Висновки до розділу	89
РОЗДІЛ 2: АДАПТАЦІЯ ТЕХНОЛОГІЇ NEURAL RADIANCE FIELDS (NeRF) ДЛЯ ЗАДАЧІ 3D РЕКОНСТРУКЦІЇ СЦЕНИ В УМОВАХ ДИНАМІЧНОГО ОСВІТДЕННЯ	92
2.1. Проблематика динамічного освітлення в задачі 3D реконструкції....	92
2.2. Методологія	93
2.2.1. Обмеження фотометричної функції втрат NeRF підходу	93
2.2.2. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Функція втрат за глибиною	94
2.2.3. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Час як додаткова вхідна змінна моделі NeRF	95
2.3. Експерименти	95

	24
2.3.1. Опис набору даних	95
2.3.2. Опис експерименту. Метрики	96
2.3.3. Результати експерименту	97
2.4. Висновки до розділу	100
РОЗДІЛ 3:РЕКОНСТРУКЦІЯ КАРТ ГЛИБИНИ ВРАХОВУЮЧИ НАПІВПРОЗОРИ ТА ВІДБИВНІ ПОВЕРХІ.....	102
3.1. Проблематика впливу напівпрозорих та відбивних поверхонь на реконструкцію карт глибини.....	102
3.2. Метод реконструкції карт глибини, що враховує напівпрозорі та відбивні поверхні.....	109
3.3. Порівняння з існуючими підходами	116
3.3.1. Відмінності від традиційного стерео-відновлення глибини	116
3.3.2. Порівняння з монокулярними мережами реконструкції карт глибини.....	116
3.3.3. Використання методу в сучасному контексті.....	118
3.4. Висновки до розділу	119
РОЗДІЛ 4:ПРОГНОЗУВАННЯ КАРТ ГЛИБИНИ ВИСОКОЇ ТОЧНОСТІ НА ПРИСТРОЯХ З ОБМЕЖЕНОЮ РОЗРЯДНІСТЮ ЗА ДОПОМОГОЮ ДВОВИМІРНИХ КРИВИХ ГІЛЬБЕРТА	121
4.1. Проблематика зниження якості реконструкції карт глибини та їх діапазону при портуванні методів на DSP/NPU.....	121
4.2. Метод.....	124
4.2.1. Прогнозування високоточних карт глибини на пристроях з низькою розрядністю	124
4.2.2. Вибір оптимальної параметричної кривої.....	127
4.2.3. Пряме та зворотне перетворення	131

	25
4.2.4. Перетворення похибки квантування	131
4.2.5. Модифікація моделі DispNet та відповідної функції втрат ..	133
4.2.6. Модифікація моделі DPT та відповідної функції втрат	134
4.3. Експерименти	135
4.3.1. Деталі реалізації	136
4.3.2. Метрики оцінки якості	136
4.3.3. Аналіз моделей у форматі W8A8	137
4.3.4. Порівняння моделей FP16, W8A16 та W8A8	140
4.3.5. Аналіз зменшення помилок квантування	144
4.3.6. Експеримент на наборі даних KITTI 2012	145
4.3.7. Вплив якості квантування на 3D реконструкцію	147
4.3.8. Експеримент з оцінювання пози людини	150
4.4. Обмеження підходу	153
4.5. Висновки до розділу	154
ВИСНОВКИ	157
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ	161

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

DP	– Доповнена реальність
2D	– (2-dimensional) Двовимірний
3D	– (3-dimensional) Тривимірний
AbsRel	– (Absolute Relative Error) Абсолютна відносна похибка
BA	– (Bundle adjustment) Алгоритм корекції пучка
CAGR	– (Compound annual growth rate) Сукупний середньорічний темп зростання
CNN	– (Convolutional Neural Network) Згорткова нейронна мережа
CPU	– (Central Processing Unit) Центральний процесор
DCT	– (Discrete Cosine Transform) Дискретне косинусне перетворення
DL	– (Deep Learning) Глибинне навчання
DLT	– (Direct Linear Transform) Пряме лінійне перетворення
DNN	– (Deep Neural Network) Глибока нейронна мережа
DSP	– (Digital Signal Processor) Цифровий сигнальний процесор
EPE	– (Endpoint Error) Похибка кінцевої точки
FPS	– (Frames Per Second) Частота кадрів / кількість кадрів за секунду
GRU	– (Gated Recurrent Unit) Керований рекурентний блок
GT	– (Ground Truth) Еталонне значення
HPE	– (Human Pose Estimation) Оцінка пози людини
LDM	– (Latent Diffusion Model) Модель латентної дифузії
LIDAR	– Light Detection and Ranging
LUT	– (Look-Up Table) Таблиця підстановки
MAE	– (Mean Absolute Error) Середня абсолютна похибка
MARE	– (Mean Absolute Relative Error) Середня абсолютна відносна похибка
MLP	– (Multilayer Perceptron) Багатошаровий перцептрон
MVS	– (Multi-View Stereo) Стерео з декількох точок спостереження
NeRF	– Neural Radiance Fields

NPU	– (Neural Processing Unit) Нейронний процесор
PTQ	– (Post Training Quantization) Квантування після навчання
QAT	– (Quantization-Aware Training) Навчання з урахуванням квантування
ReLU	– (Rectified Linear Unit) Випрямлений лінійний вузол, або випрямляч
RGB	– (Red, Green, Blue) Кольорова модель зображення
RMSE	– (Root Mean Square Error) Середньоквадратична похибка
SD	– (Standard Deviation) Стандартне відхилення
SDF	– (Sign distance functions) Функція знакової відстані
SfM	– (Structure from Motion) Структура з руху
SIFT	– (Scale Invariant Feature Transform) Масштабно-незмінне перетворення ознак
SSIM	– (Structural Similarity Index Measure) Індекс структурної подібності
ToF	– Time-of-flight (тип камери)
TPU	– (Tensor Processing Unit) Тензорний процесор
TSDF	– (Truncated Signed Distance Function) Усічена знакова функція відстані
VAE	– (Variational Autoencoder) Варіаційний автоенкодер
ViT	– (Vision Transformer) Візуальний трансформер
W8A16	– 8-бітні ваги та 16-бітні активації
W8A8	– 8-бітні ваги та 8-бітні активації

ВСТУП

Актуальність роботи. Методи машинного та глибинного навчання, а також підходи штучного інтелекту є гнучкими інструментами, що знаходять застосування в широкому спектрі технологічних напрямів: від медицини [1, 2] та безпеки [3] до інтерфейсів мозок—комп'ютер [4], дистанційного зондування Землі та геопросторового інтелекту [5–7]. Важливу роль ці інструменти відіграють у задачах комп'ютерного зору, де забезпечують автоматизований аналіз візуальних даних і високу точність розпізнавання та інтерпретації сцен. До типових прикладів належать оптичне розпізнавання тексту, оцінювання глибини, виявлення та класифікація об'єктів, сегментація зображень, відстеження руху, оцінювання поз, а також 3D реконструкція та розуміння сцени [8–16]. Сукупність цих можливостей робить технології машинного та глибинного навчання ключовими також для трендового напрямку доповненої та віртуальної реальності, де потрібне надійне сприйняття простору та взаємодія з цифровими об'єктами в реальному часі.

В останні десятиліття спостерігається надзвичайно швидкий розвиток технологій доповненої реальності (ДР), які відкривають новий етап взаємодії людини з інформацією та оточуючим середовищем. Сценарії використання цих технологій стрімко еволюціонують, переходячи з мобільних на самостійні-носимі пристрої. Продукти, які дозволяють споживати сценарії ДР, виходять зі стадії обмежених та дорогих платформ розробки і набувають масовості. Розмір світового ринку ДР на 2023 рік складав 43,4 млрд \$. При прогнозованому сукупному середньорічному темпі зростання (CAGR) у 34,3% він має перевищити 616 млрд \$ у 2032 році [17].

Сучасні тренди у галузі доповненої реальності визначаються не лише зростанням її популярності, але й постійним удосконаленням технологічного арсеналу. Все більше компаній і дослідницьких груп зосереджують свою увагу на розробці інноваційних методів та рішень спрямованих на підвищення реалістичності взаємодії об'єктів ДР з реальним світом та користувача з об'єктами ДР. Інтенсивність розвитку цього напрямку не тільки збільшує

можливості технологій ДР, але і відкриває нові перспективи для їхнього впровадження у різноманітних сферах життя, від освіти та розваг до промисловості та медицини.

Одні з ключових напрямків цього еволюційного процесу є реконструкція карт глибини та 3D реконструкція оточуючого середовища і його об'єктів, що стає невід'ємною складовою при досягненні максимальної імерсії та реалізації потенціалу ДР. Мета 3D реконструкції полягає у відтворенні тривимірної моделі оточення на основі доступних даних: дво- та тривимірні зображення, внутрішні та зовнішні параметри камери, дані з IMU сенсорів, або систем позиціонування і т.д. 3D реконструкція є своєрідним каркасом для доповненої реальності [18]. Якість отриманої моделі суттєво впливає на реалістичність сприйняття сценаріїв доповненої реальності користувачем. Справжній виклик полягає в тому, як забезпечити точність, достовірність та ефективність тривимірної моделі. В цьому контексті, доволі часто, важливу роль відіграє оцінка карт глибини. Методи реконструкції карт глибини приймають на вхід монокулярне зображення, стереозображення, або їх послідовність, та повертають результуюче зображення, де кожний піксель представляє собою відносну, або абсолютну відстань від камери до відповідної точки простору. Оцінка глибини дозволяє системам реконструкції ефективно розміщувати об'єкти в просторі, враховуючи їхнє точне положення та взаємодію. Це стає критичним у віртуальних і доповнених середовищах, де недостовірність 3D реконструкції може призвести до неправильної взаємодії з оточенням та втрати реалістичності.

Об'єктом дослідження є процеси тривимірної реконструкції середовища для систем доповненої реальності.

Предметом дослідження є математичні моделі, алгоритми та методи обробки зображень і даних з різних джерел для отримання високоточної 3D реконструкції в умовах динамічного освітлення, складних поверхонь та обмежених обчислювальних ресурсів.

Мета та задачі дослідження. Метою даного дослідження є підвищення якості, стійкості та енергоефективності методів 3D реконструкції середовища,

що виконуються на кінцевому пристрої користувача, для забезпечення реалістичної та надійної взаємодії користувача з віртуальними об'єктами в системах доповненої реальності.

Для досягнення поставленої мети необхідно вирішити наступні задачі:

1. Виконати аналіз класичних та сучасних методів реконструкції карт глибини та 3D реконструкції середовища, визначити їхні обмеження для задач доповненої реальності на кінцевих пристроях користувача.
2. З метою врахування динамічних сцен та розширення існуючих наборів даних для задачі 3D реконструкції середовища, адаптувати технологію Neural Radiance Fields (NeRF) до умов динамічного освітлення.
3. Запропонувати метод обробки напівпрозорих та відбивних поверхонь для покращення реконструкції складних об'єктів.
4. Розробити метод ефективного прогнозування карт глибини високої точності з широким діапазоном глибини на кінцевих пристроях користувача та апаратних прискорювачах з обмеженою розрядністю.

Методи дослідження. Дослідження ґрунтуються на методах комп'ютерного зору та обробки зображень, машинного і глибинного навчання, математичного моделювання та чисельних експериментів, статистичної обробки результатів, а також методах квантування та оптимізації нейронних мереж.

Наукова новизна отриманих результатів:

1. Удосконалено метод Neural Radiance Fields (NeRF) до умов динамічного освітлення шляхом модифікації функцій втрат та введення часової змінної, що дозволило покращити якість реконструкції сцен з динамічним освітленням та розширювати існуючі набори даних для складних сцен.
2. Вперше розроблено та запатентовано метод реконструкції глибини, який враховує напівпрозорі та відбивні поверхні, зберігає окремо значення глибини до самої площини та до відбитого/перекритого

об'єкту, що дозволяє збільшити точність реконструкції складних сцен, які містять не дифузні поверхні.

3. Вперше сформульовано та розроблено метод для прогнозування карт глибини з представленням виходу моделі у вигляді компонент двовимірної кривої Гільберта [19], що дозволило для квантованих моделей розширити діапазон глибини, підвищити її точність та покращити енергоефективність реконструкції сцени на апаратних прискорювачах з обмеженою розрядністю (DSP/NPU) кінцевих пристроїв користувача.

Практичне значення отриманих результатів та їх застосування.

Результати присутні в даному дослідженні були використані в рамках комерційних та науково-дослідницьких проектів ТОВ «Самсунг РнД Інститут Україна». Запропоновані методи знайшли застосування у галузі візуального інтелекту при вирішенні завдань 3D реконструкції сцени для доповненої реальності та створення/редагування просторового контенту.

1. Метод призначений для реконструкції карт глибини з урахуванням напівпрозорих та відбивних поверхонь [20], що орієнтований на підвищення якості відтворення складних об'єктів, зокрема при наявності динамічного освітлення захищено патентом (US20240144503A1). Отриманий патент розширює патентне портфоліо компанії у відповідному технологічному домені.
2. Метод прогнозування високоточних карт глибини з широким діапазоном на пристроях з обмеженою розрядністю, що ґрунтується на використанні двовимірних кривих Гільберта, дозволяє: зменшити похибку квантування моделі у 4,6 рази, що підвищує якість реконструкції карт глибини на DSP-пристроях; розширити діапазон відстаней за рахунок збільшення ефективної розрядності з 8-біт до 10-біт; зменшити час виконання та енергоспоживання квантованої моделі у 1,5 раза порівняно з оригінальною моделлю за умови збереження або покращення якості прогнозування карт глибини.

3. Розглянуті та запропоновані методи 3D реконструкції сцени на кінцевих пристроях користувача з обмеженими обчислювальними ресурсами було використано при розробці комерційних проектів, що спрямовані на сценарії створення/редагування просторового контенту для флагманської моделі смартфона Samsung Galaxy S25.

Також, представлення виходу моделі на основі кривої Гільберта, що запропонована у методі прогнозування карт глибини високої точності з широким діапазоном на пристроях з обмеженою розрядністю за допомогою двовимірних кривих Гільберта може бути застосована для покращення якості та ефективності і інших задач комп'ютерного зору на основі методів машинного навчання, що виконуються на апаратних прискорювачах з низькою арифметикою. Проведені додаткові експерименти для оцінки впливу запропонованого підходу на задачу оцінки пози людини (Human Pose Estimation, HPE) демонструють зменшення похибки квантування на DSP в 2.69 рази [21].

Особистий внесок здобувача. Усі основні результати дисертаційного дослідження, представлені до захисту, одержані автором самостійно.

В публікаціях у співавторстві, здобувачеві належать такі результати:

У роботі [22] здобувачем: сформульовано проблему та мотивацію роботи; проведено огляд літератури в області використання Neural Radiance Fields для задач 3D реконструкції та синтезу наборів даних; підготовлено науковий контекст задачі; підготовлено дані для експериментів (відбір сцен/кадрів, попередня обробка, формування навчальних та валідаційних підмножин); здійснено підбір і валідацію гіперпараметрів навчання; запропоновано і проведено експерименти та порівняльний аналіз отриманих результатів, що підтвердило перевагу запропонованого підходу над оригінальною моделлю для задач 3D реконструкції сцени в умовах динамічного освітлення; написання тексту статті; проведено роботу із зовнішніми рецензентами.

У роботі [20] здобувачем: проведено аналіз існуючих робіт та винаходів в області реконструкції глибини враховуючи не дифузні поверхні; запропоновано ідею методу реконструкції глибини, який враховує напівпрозорі та відбивні

поверхні; запропоновано та проведено ряд експериментів для підтвердження доцільності методу реконструкції глибини, який враховує напівпрозорі та відбивні поверхні, зберігає окремо значення глибини до самої площини та до відбитого/перекритого об'єкту, що дозволяє збільшити точність реконструкції складних сцен; оформлено відповідні розділи патентної заявки; проведена робота з патентним бюро по узгодженню деталей запропонованого методу перед публікацією патенту.

У роботі [21] здобувачем: проведено аналіз робіт попередніх дослідників орієнтованих на підвищення точності та енергоефективності реконструкції сцени; удосконалено оригінальну ідею та прийнято участь в розробці методу її технічної реалізації; проведено аналіз та запропоновано оптимальну параметричну криву та її порядок, що дозволило реалізувати метод прогнозування карт глибини з представленням виходу моделі у вигляді компонент двовимірної кривої Гільберта; запропоновано та реалізовано зворотне перетворення результатів виходу моделі та алгоритм постобробки, що сприяло розширенню діапазону глибини для квантованої моделі при її виконанні на прискорювачах з обмеженою розрядністю (DSP/NPU), підвищенню її точності та енергоефективності; обрано метрики оцінки якості запропонованого підходу; проведено ряд експериментів, що підтверджують переваги запропонованого підходу; підготовлено чорновий варіант статті.

Апробація матеріалів дисертації. Результати та основні положення роботи подавалися та обговорювалися на:

- Kolodiazhna O., Savin V., Uss M., Kussul N., 3D Scene Reconstruction with Neural Radiance Fields (NeRF) considering dynamic illumination conditions, *Proceedings of International Conference on Applied Innovation in IT 2023*, 2023, Volume 11, Issue 1, pp. 233-238. ISSN: 2199-8876, DOI: 10.25673/101943. [Scopus]
- Uss M., Iermolenko R., Shashko O., Kolodiazhna O., Safonov I., Savin V., Yeo Y, Ji S., Jeong J., Predicting High-precision Depth on Low-Precision Devices Using 2D Hilbert Curves, *Proceedings of the 42nd*

International Conference on Machine Learning, PMLR, 2025, vol. 267, pp. 60635 – 60656, ISSN: 2640-3498. [Scopus]

Публікації. За результатами досліджень опубліковано:

- Статтю у науковому фаховому виданні України [22], що включено до списку міжнародних наукометричних баз Scopus та Web of Science з квантилем Q3 та публікується українською і англійською мовами.
- Міжнародний патент на винахід безпосередньо за напрямком дослідження [20].
- Одноосібний розділ у монографії [18].

Окрім того, на момент написання дисертаційної роботи, здобувач має 6 додаткових опублікованих патентів за суміжними напрямками досліджень.

Структура та обсяг дисертації. Дисертаційна робота складається із анотації (українською та англійською мовами), списку публікацій здобувача, змісту, переліку умовних позначень, вступу, чотирьох розділів, загальних висновків та списку використаних джерел. Робота містить: 133 сторінки основного тексту; 66 рисунків; 8 таблиць. Список використаних джерел містить 178 найменувань і займає 20 сторінок. Загальний обсяг дисертаційної роботи – 180 сторінок.

РОЗДІЛ 1: АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ТА ПІДХОДІВ ЗД РЕКОНСТРУКЦІЇ СЦЕНИ

Методи машинного навчання сьогодні є універсальним і гнучким інструментом для розв'язання широкого спектра прикладних задач. Їх ключова перевага полягає у здатності виявляти закономірності в даних, будувати прогностичні моделі та адаптуватися до різних типів вхідної інформації без потреби у повному явному програмуванні всіх правил обробки. Завдяки цьому машинне навчання знайшло широке застосування в багатьох напрямках сучасної науки й техніки.

У галузі комп'ютерного зору методи машинного навчання використовуються для класифікації зображень, детекції та сегментації об'єктів [5–7], відстеження рухомих об'єктів, оцінювання пози та жестів людини [12, 13], розпізнавання облич, аналізу медичних зображень [1, 2], відновлення карт глибини, побудови тривимірних моделей сцени [9–11, 14, 15], а також для розпізнавання тексту на зображеннях і в документах [8]. В задачах аналізу часових рядів ці методи застосовуються для прогнозування значень у часі, виявлення аномалій, аналізу сенсорних сигналів, фінансового прогнозування, моделювання попиту, обробки біомедичних сигналів та аналізу послідовностей. Окремо слід зазначити задачі обробки природної мови, зокрема розпізнавання та аналіз натуральної мови, автоматичний переклад, класифікацію текстів, аналіз тональності, побудову діалогових систем і пошук інформації. Активний розвиток методів машинного навчання в цих та інших прикладних областях є одним із ключових чинників інтенсивного розвитку сфери штучного інтелекту загалом.

Лінійна регресія [23] є одним із базових методів машинного навчання, що застосовується для розв'язання задач прогнозування неперервних величин. Її ідея полягає у побудові лінійної залежності між набором вхідних ознак і цільовою змінною. Попри відносну простоту, цей метод залишається важливим завдяки зрозумілості, інтерпретованості та низькій обчислювальній складності. Лінійна регресія часто використовується як базова модель для порівняння з більш складними підходами.

Одношаровий перцептрон [24] є однією з найперших нейромережових моделей, призначених переважно для задач бінарної класифікації. Він реалізує просте лінійне правило прийняття рішення на основі зваженої суми вхідних ознак. Основним обмеженням цієї моделі є здатність коректно працювати лише для лінійно роздільних даних. Водночас саме перцептрон став основою подальшого розвитку нейронних мереж.

Багатошаровий перцептрон [25] є узагальненням одношарового перцептрона та містить один або декілька прихованих шарів нейронів. Наявність нелінійних функцій активації дає змогу моделі апроксимувати складні нелінійні залежності між вхідними даними та виходом. Багатошарові перцептрони застосовуються у задачах класифікації, регресії та апроксимації функцій. Саме поява ефективних алгоритмів навчання багатошарових мереж стала важливим етапом у становленні сучасного машинного навчання.

Дерева рішень [26] належать до методів, які виконують послідовне розбиття простору ознак на області, що відповідають різним рішенням або прогнозам. Вони є зручними з точки зору інтерпретації, оскільки результат моделювання можна подати у вигляді зрозумілої послідовності умов. Проте окреме дерево рішень може бути нестійким до змін у даних і схильним до перенавчання.

Для зменшення цього недоліку застосовується **метод випадкового лісу**, який об'єднує велику кількість дерев рішень, побудованих на різних підмножинах даних та ознак. Підсумкове рішення формується шляхом агрегування результатів окремих дерев. Такий підхід зазвичай забезпечує вищу точність і кращу узагальнювальну здатність порівняно з одним деревом.

Метод опорних векторів (Support Vector Machine, SVM) [27] є потужним інструментом для задач класифікації та регресії. Його основна ідея полягає у побудові розділяючої гіперплощини з максимальним зазором між класами. Завдяки використанню ядерних функцій цей метод може ефективно працювати і в випадках, коли дані не є лінійно роздільними у початковому просторі ознак.

SVM добре зарекомендував себе на задачах середньої розмірності та при обмеженому обсязі навчальних даних.

Кластеризація [28] належить до методів навчання без учителя та використовується для виявлення прихованої структури в даних. Її метою є поділ множини об'єктів на групи таким чином, щоб об'єкти всередині однієї групи були більш подібними між собою, ніж до об'єктів інших груп. Методи кластеризації широко застосовуються для попереднього аналізу даних, сегментації користувачів, групування документів, аналізу зображень та виявлення аномалій. До найпоширеніших підходів належать метод k-середніх, ієрархічна кластеризація та метод DBSCAN.

Глибинне навчання [29] є сучасним напрямом машинного навчання, що базується на використанні багатошарових нейронних мереж зі складною ієрархічною структурою. Його ключова перевага полягає у здатності автоматично формувати інформативні ознаки без необхідності ручного проєктування великої кількості дескрипторів. Саме це дозволило досягти суттєвого прогресу в задачах комп'ютерного зору, розпізнавання мовлення, обробки природної мови, аналізу часових рядів та робототехніки.

До найважливіших архітектур глибинного навчання належать згорткові нейронні мережі, які особливо ефективні для обробки зображень, рекурентні нейронні мережі та їх модифікації для аналізу послідовностей, а також трансформерні архітектури, що стали основою сучасних рішень у сфері аналізу тексту, мультимодальних даних і генеративних моделей. Глибинне навчання стало технологічною основою багатьох сучасних інтелектуальних систем.

Таким чином, машинне навчання охоплює широкий спектр методів – від простих лінійних моделей до складних глибинних нейронних мереж – і забезпечує ефективний інструментарій для розв'язання різноманітних прикладних задач. Розвиток цих методів безпосередньо сприяє прогресу в галузі штучного інтелекту та відкриває нові можливості для аналізу даних, автоматизації прийняття рішень і побудови інтелектуальних систем.

У цьому контексті дане дисертаційне дослідження позиціонується як розвиток і покращення методів машинного навчання, зокрема методів комп'ютерного зору, орієнтованих на розв'язання задачі 3D реконструкції для доповненої реальності. Запропоновані в роботі підходи спрямовані на підвищення точності, стійкості та практичної придатності моделей реконструкції просторової структури сцени в умовах, характерних для реальних прикладних систем.

Доповнена реальність представляє собою технологію, яка дозволяє в режимі реального часу відображати штучний 2D/3D контент та інформацію поверх або поруч фізичних об'єктів реального світу. Доповнені об'єкти природньо інтегруються у фізичне середовище [18] за рахунок широкого алгоритмічного стеку технологій комп'ютерного зору, який можна умовно розподілити на наступні категорії: калібровка сенсорів пристрою споживання ДР (заводська, статична та динамічна калібровка) [30–32], відстеження позиції пристрою ДР в 6-ти ступенях свободи та супроводження об'єктів [12, 13], розуміння сцени (аналіз карти глибини та 3D реконструкцію, розпізнавання об'єктів та семантичне розуміння сцени, виявлення джерел освітлення та ін.) [14–16, 33, 34], взаємодія між користувачем та пристроєм ДР (голосова взаємодія, відстеження напряму погляду, розпізнавання жестів рук, мультимодальна взаємодія, та ін.) [35–40]. Швидкий розвиток цього набору технологій дає змогу підвищити реалістичність сценаріїв доповненої реальності та сприймати штучні об'єкти як частину оточення.

На відміну від віртуальної реальності, де користувач занурюється у повністю штучне середовище, доповнена реальність надає можливість взаємодії зі штучним контентом не втрачаючи зв'язок з оточенням. Доповнені 2D/3D об'єкти прив'язуються до певних фізичних об'єктів, взаємодіють з ними та з іншими предметами сцени за рахунок оклюзій та колізій (Рис. 1.1). Реалістичність таких взаємодій залежить від якості «каркасу», на який вони накладаються. Цим каркасом виступає 3D реконструкція середовища.

Останнім часом доповнена реальність широко застосовується в різноманітних галузях, що включають: освіту, медицину, інженерію, зв'язок та віддалену підтримку, сферу розваг та багатьох інших напрямків [41–46]. Також, доповнена реальність знаходить і військове застосування.

На сьогоднішній день найдоступнішим пристроєм, який дозволяє споживати сценарії ДР виступає мобільний телефон. Зображення, отримане з камери, ретранслюється на екран та доповнюється штучними об'єктами. Основні недоліки такої експлуатації: телефон треба тримати у руках, користувач спостерігає лише невеличку область сцени, відсутність реалістичного сприйняття середовища.



Рис. 1.1. Приклад сценарію ДР з урахуванням оклюзій та колізій

Стрімкий розвиток носимих пристроїв ДР націлений на позбавлення вищезгаданих недоліків та на популяризацію технології. За формфактором носимі пристрої ДР можна розділити на:

- Формфактор шолому віртуальної реальності (Рис. 1.2, а). Пара камер ретранслює зображення середовища, яке доповнюється, на пару екранів високої роздільної здатності та частоти, що знаходяться в середині шолому перед очима користувача. Додатковою перевагою цієї конструкції є можливість споживання сценаріїв віртуальної реальності. Основні недоліки: великий розмір та вага, що впливають на ергономічність використання.

- Формфактор окулярів (Рис. 1.2, б). Зображення з камер передаються на прозорі дисплеї, наприклад Waveguide [47], Transparent MICRO LED [48], або інші. Користувач спостерігає оточуюче середовище крізь них. Доповнений контент відображається на дисплеях поверх реальних об'єктів. Основний недолік: мініатюрний розмір впливає на можливість розміщення апаратних засобів та батарей високої ємності.



а



б

Рис. 1.2. Приклад носимих пристроїв ДР: формфактор шолому віртуальної реальності (а) – Meta Quest Pro [49], формфактор окулярів (б) – XREAL Air 2 Pro [50]

За критерієм розміщення апаратно-обчислювальних потужностей носимі пристрої ДР можна розділити на:

- Самостійні (Рис. 1.2). Апаратно-обчислювальні потужності розміщені на самому пристрої.
- Пов'язані (Рис. 1.3). Апаратно-обчислювальні потужності частково або повністю розміщені на пристрої-компаньйоні. Пристрої можуть з'єднуватись як провідним, так і безпроводним шляхом



Рис. 1.3. Приклад пов'язаних пристроїв ДР (Magic Leap 2 [51])

Сучасна тенденція розвитку пристроїв ДР спрямована на мініатюризацію з метою перетворити громіздкий гаджет на ергономічний повсякденний аксесуар

та помічник. Однак, зменшення розмірів змушує використовувати батареї невеликої ємності. Це, в свою чергу, впливає на можливість встановити або активно використовувати сенсори з підвищеним енергоспоживанням, такі як: ToF (Time of Flight) [52], LIDAR [53] та сенсори структурного світла [54]. Отже, доволі часто, реконструкцію карт глибини та/або 3D реконструкцію сцени доводиться проводити на основі даних з монокулярної, або стерео камери, що є доцільніше з точки зору оптимізації енергоспоживання. Тому, в даній роботі, передусім, розглядаються методи та підходи орієнтовані на реконструкцію карт глибини та 3D реконструкцію сцени саме на основі даних з монокулярної та стерео камери.

1.1. Методи реконструкції карти глибини по 2D даним

Важливою складовою 3D реконструкції є розуміння карти глибини (відстані до кожної точки спостереження). Людині зазвичай легко сприймати інформацію про тривимірну структуру об'єкта або сцени та оцінювати відстань до об'єктів [18]. Але визначення карти глибини по зображенню або їх серії є складною задачею комп'ютерного зору, оскільки під час зйомки відбувається проєкція сцени на площину, що призводить до втрати третього виміру.

Методи реконструкції карти глибини з 2D даних можна умовно поділити на два класи в залежності від кількості вхідних зображень:

- методи, що використовують одне вхідне нерухоме зображення та спираються на монокулярні ознаки глибини;
- методи, що базуються на аналізі двох і більше зображень та оперують багатоокулярними (multi-ocular) ознаками глибини.

У другому випадку два або більше вхідних зображень можуть бути зроблені кількома фіксованими камерами з різних кутів огляду або однією камерою, що рухається, у різні проміжки часу.

Табл. 1.1. Основні ознаки глибини, що використовуються методами реконструкції глибини по 2D даним

Кількість вхідних зображень	Ознаки глибини
Одне зображення	Лінійна перспектива
	Атмосферне розсіювання
	Розуміння форми по затіненню
Два і більше зображень	Бінокулярна диспаратність
	Паралакс руху
	Розмиття зображення
	Силует
	Структура з руху

1.1.1. Лінійна перспектива

В основі лінійної перспективи лежить ідея, що паралельні лінії, такі як дороги або стежки, збігаються вдалині. Точки перетину цих ліній менш помітні, ніж точки ліній наближені до спостерігача. Підхід, запропонований в [55], працює для зображень, що містять поверхні з жорсткою геометрією. Точка з найбільшою кількістю перетинів у певному районі вважається точкою зникнення. Основні лінії поблизу точки зникнення позначаються як лінії зникнення (Рис. 1.4).

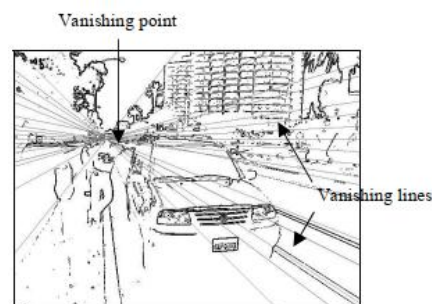


Рис. 1.4. Детектування ліній і точки зникнення [55]

Між кожною парою сусідніх ліній зникнення призначається набір градієнтних площин, кожна з яких відповідає окремому рівню глибини. Пікселі ближче до точок зникнення отримують більше значення глибини, і щільність градієнтних площин вища (Рис. 1.5).

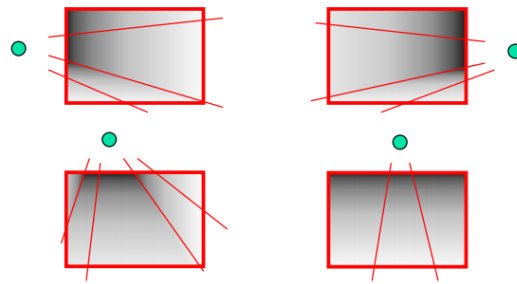


Рис. 1.5. Приклади евристичних правил для створення градієнтних площин глибини, де зелене коло – точка зникнення [55]

1.1.2. Атмосферне розсіювання

Підхід атмосферного розсіювання ґрунтується на тому, що потужність і напрям світла змінюються, коли світло проходить через атмосферу через наявність у ній дрібних частинок. Об'єкти, які знаходяться ближче до камери, виглядають чіткіше, тоді як більш віддалені об'єкти – розмиті. У [56] представлено аналіз цього перетворення, заснований на фізичній моделі розсіювання лорда Релея 1871 року. Їхній алгоритм підходить для оцінки глибини зображень на відкритому повітрі, які містять частину неба (Рис. 1.6).

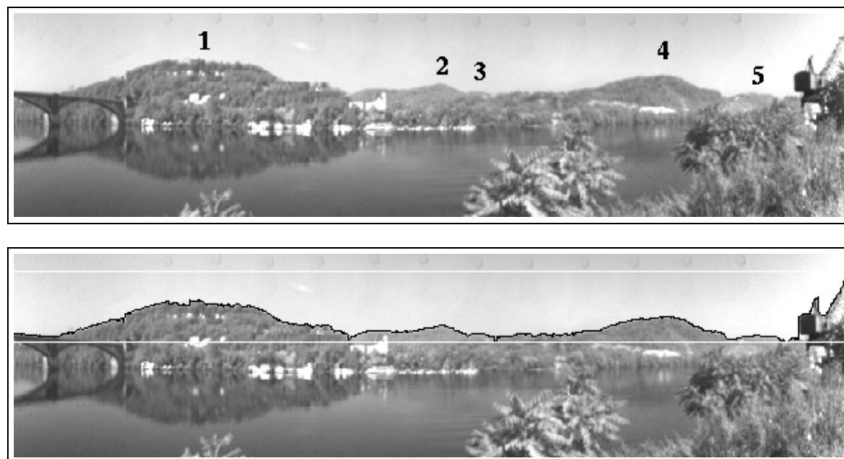


Рис. 1.6. Приклад детектування горизонту та зон атмосферного розсіювання [56]

1.1.3. Розуміння глибини по затіненню

Метод розуміння форми по затіненню дозволяє визначити нормаль поверхні об'єкта, спостерігаючи за відбивною здатністю світла на цьому об'єкті. Кількість світла, яка відбивається від поверхні об'єкта, залежить від його орієнтації. Вперше цю ідею представив Woodham у 1980 році [57]. Підхід

розуміння форми по затіненню (shape from shading), використовується для аналізу одного вхідного зображення та був представлений В. К. Horn в 1989 році [57]. Фотометричний стереоаналіз відтоді був узагальнений для багатьох інших ситуацій, таких як, наприклад не Ламбертові поверхні. Процес реконструкції карти глибини по затіненню на основі зображень світлового поля представлений на Рис. 1.7.

По декільком зображенням об'єкта при різному освітленні можна провести оцінку векторів нормалей у кожному пікселі [57].

Метод потребує специфічного технічного обладнання.

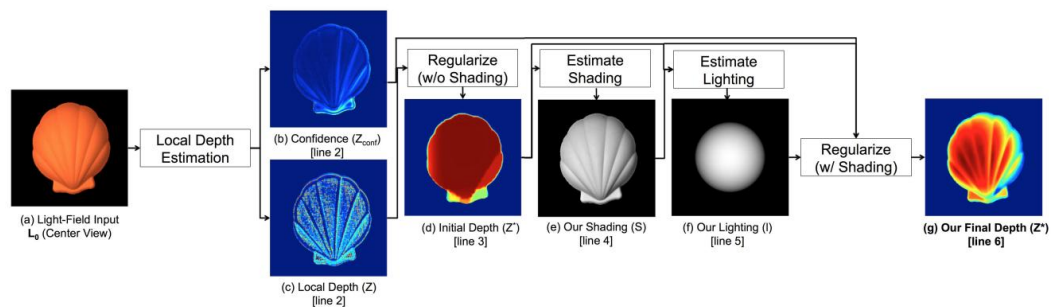


Рис. 1.7. Ілюстрація підходу відтворення глибини по затіненню на основі зображень світлового поля [57]

1.1.4. Бінокулярна диспаратність

Використовуючи два зображення однієї сцени, зроблені одночасно з дещо різних точок спостереження, можна відновити глибину точки, що присутня на обох зображеннях. Спочатку знаходиться відповідний набір точок на обох зображеннях. Потім, для кожного набору точок використовується метод триангуляції для визначення глибини відповідної фізичної точки, що була зпроектована на пару зображень [58, 59].

На Рис. 1.8 проілюстрована система стереоскопічного зору для якої проводиться обрахунок бінокулярної диспаратності, де:

- P – точка у просторі, що належить фізичному об'єкту;
- C_l та C_r – ліва та права камера стереосистеми;
- P_l та P_r – проєкції точки на матриці лівої (C_l) та правої камер (C_r). P_l та P_r знаходяться на епіполярній лінії;

- x_l та x_r – відповідні зсуви проекції точки від початку системи координат зображення (x координати точок P_l та P_r);
- Z – значення глибини (відстань до P);
- f – фокусна відстань;
- B – відстань між камерами стереопари (baseline).

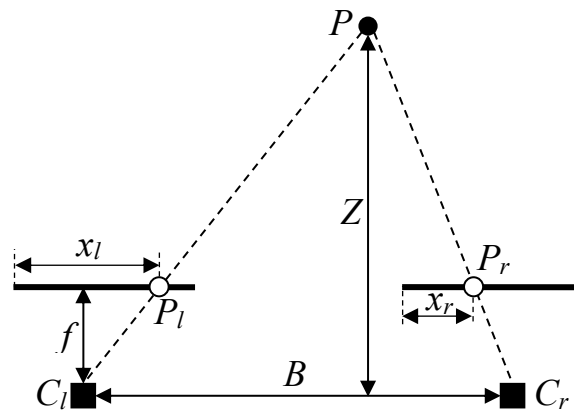


Рис. 1.8. Ілюстрація біноклярного диспаритету

Значення глибини Z можна розрахувати спираючись на принципи триангуляції, що застосовуються в системах стереоскопічного зору:

$$Z = f \frac{B}{D} \quad (1.1)$$

де D – значення диспаритету, який розраховується як різниця між x_l та x_r .

1.1.5. Паралакс руху

Відносний рух між камерою і сценою надає важливі ознаки для сприйняття глибини. Об'єкти, які знаходяться ближче до камери, рухаються швидше, ніж ті, що знаходяться далі. Відновлення тривимірних структур називається реконструкцією структури з руху. Рух можна розглядати як форму диспаратності протягом часу, що представлено поняттям поля руху. Поле руху – це двовимірні вектори швидкості точок зображення та спостережуваної сцени. Основні припущення для структури з руху полягають у тому, що об'єкти не деформуються і їхні рухи є лінійними. Ці властивості було використано у кількох методах, таких як "стереоскопія погойдування (wiggle stereoscopy)" [60], де паралакс руху

використовується як представлення для стереоскопічних зображень, або "стереоскопія прокручування (parallax scrolling)" [61], що широко використовується в комп'ютерній графіці, де шляхом руху переднього і заднього планів з різною швидкістю викликається відчуття глибини. Вплив цієї ознаки глибини є відносно сильним в порівнянні з іншими монокулярними ознаками глибини, а також у порівнянні з бінокулярною диспаратністю.

1.1.6. Розмиття зображення

Методи визначення глибини за розмиттям зображення дозволяють реконструювати карту глибини на основі ступеня розмиття, присутнього на зображеннях. У системі з тонкою лінзою об'єкти, які знаходяться у фокусі, відображаються чітко, тоді як об'єкти на інших відстанях розфокусовані, тобто розмиті. На Рис. 1.9 показана модель тонкої лінзи з реальною точкою P , що знаходиться за межами фокусної відстані лінзи. Відповідна проекція на площину зображення є кругова розмита пляма зі сталою яскравістю, центрованою у точці P'' та радіусом розмиття σ .

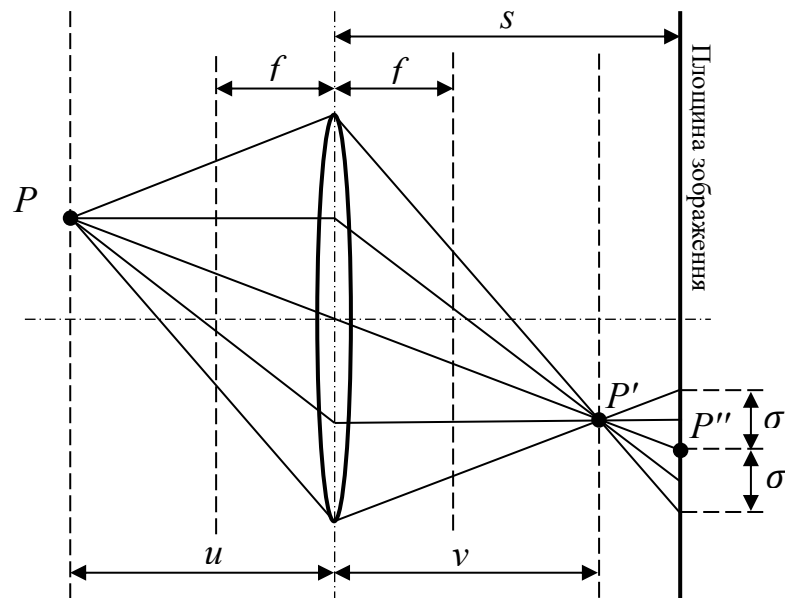


Рис. 1.9. Ілюстрація моделі тонкої лінзи

Для того щоб оцінити глибину u , нам необхідні наступні рівняння. Основне рівняння, що описує співвідношення між u , v та f для тонких лінз:

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \quad (1.2)$$

У [62] представлено виведення співвідношення між відстанню u та розмиттям σ у рівнянні:

$$u = \begin{cases} \frac{fs}{s - f - kf\sigma}, & u > v \\ \frac{fs}{s - f + kf\sigma}, & u < v \end{cases}, \quad (1.3)$$

де:

- u – глибина;
- v – відстань між лінзою та точкою ідеального фокусу;
- s – відстань між лінзою та площиною зображення;
- f – фокусна відстань лінзи;
- k – константа, визначена оптичною системою;
- σ – радіус розмиття.

Задача обчислення глибини u перетворюється на задачу оцінки параметрів камери (s , f та k) та параметра розмиття σ . Параметри камери можуть бути отримані шляхом її калібрування. Глибину u можна обчислити з рівняння (1.3), якщо відомий параметр розмиття σ .

1.2. Класичні методи 3D реконструкції середовища

3D реконструкцію в режимі реального часу можна визначити як процес, при якому відбувається відновлення віртуальної тривимірної моделі сцени, або об'єкта на ній, по зображенням з камери. Окрім даних з камери необхідними є параметри камери та її положення під час зйомки. Зазвичай ця інформація відома, або обчислюється на послідовності зображень.

3D реконструкцію сцени можна зробити, використовуючи або одне зображення, або кілька знімків, зроблених з різних позицій камери. Високу популярність набрала 3D реконструкція на основі множини зображень, де використовуються такі базові підходи як стерео-зір, структура з руху (SfM) та

стерео підходи на основі даних з декількох точок спостереження (Multi-View Stereo – MVS). Активний розвиток глибинного навчання дозволив проводити реконструкцію середовища навіть по одному зображенню [18].

1.2.1. Відтворення форми за силуетом

Силует об'єкта на зображенні відноситься до контуру, який відділяє об'єкт від фону. Методи визначення форми за силуетом вимагають кілька видів сцени, знятих камерами з різних точок спостереження. Такий процес разом із правильною текстуризацією створює повну 3D модель об'єктів у сцені. Метод відтворення форми за силуетом потребує точної калібровки камери.

Процедура 3D реконструкції, що базується на аналізі силуетів називається відтворення форми за силуетом (shape-from-silhouette) описана в [63]. Для кожного зображення силует цільового об'єкту сегментується за допомогою віднімання фону. Отримані силуети проєктуються назад у загальний 3D простір з проєкційними центрами, рівними положенням камер. Зворотне проєктування силуету створює конусоподібний об'єм. Перетин усіх конусів утворює візуальну оболонку цільового 3D об'єкта, який часто зберігається у воксельному вигляді.

На Рис. 1.10 C позначено куб, який є прикладом 3D об'єкта; S позначає двовимірний екран; P_A та P_B – точки спостереження в 3D просторі; D_A – двовимірний багатокутник на екрані, який є силуетом куба; V_A та V_B – конусоподібний об'єм, зворотно спроектований з точок спостереження P_A та P_B .

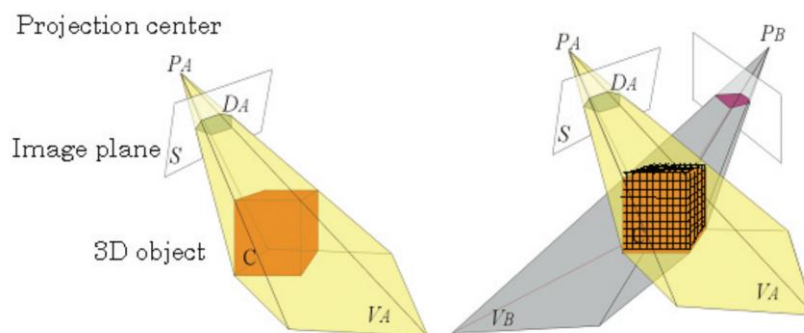


Рис. 1.10. Перетин об'ємів силуетів [63]

1.2.2. Структура з руху

Метод відновлення структури з руху (Structure from Motion – SfM) використовує серію двовимірних зображень сцени або об'єкта для відновлення їх розрідженого об'єму точок та положень камери (позиція та орієнтація).

SfM базується на тих же принципах, що й стереоскопічна фотограмметрія. У стереофотограмметрії використовується триангуляція для обчислення відносних тривимірних позицій (x, y, z) об'єктів з пар стереозображень. Такі методи вимагають використання стерео камер та відповідного програмного забезпечення. На відміну від них, стандартні (монокулярні) камери добре підходять для методів SfM. Зображення часто знімають під час руху однієї або декількох камери з різних точок спостереження [18].

Метод відновлення структури з руху (SfM) здатен проводити реконструкцію 3D структури на базі знайденого розрідженого об'єму точок, що отриманий з використання серії двовимірних зображень сцени або об'єкта. Для створення 3D реконструкції із застосуванням SfM необхідно мати багато зображень області або об'єкта з високим ступенем перекриття, зроблених з різних точок спостереження. SfM проілюстровано на Рис. 1.11. Алгоритм включає три основні етапи:

1. Співставлення відповідних ознак і вимірювання відстаней між ними на площині зображення камери d та d' . Алгоритм Scale Invariant Feature Transform (SIFT) [64] дозволяє порівнювати відповідні ознаки навіть за великих варіацій у масштабі та куті огляду, а також за умов часткового перекриття і зміни освітлення.
2. При наявності відповідних місць розташування декількох точок на двох або більше зображеннях, зазвичай існує лише одне математичне рішення для визначення місця, де були зроблені фотографії. Тому можна обчислити позиції камер (x, y, z) , (x', y', z') , орієнтації i та i' , фокусні відстані f та f' , і відносні позиції відповідних ознак b та h в одному кроці, який називається bundle adjustment (BA). Звідси походить термін

"структура з руху". Структура сцени включає всі ці параметри, а рух відноситься до руху камери.

3. Наступним кроком є визначення щільного об'єму точок і 3D поверхні, використовуючи параметри камери і точки SfM.

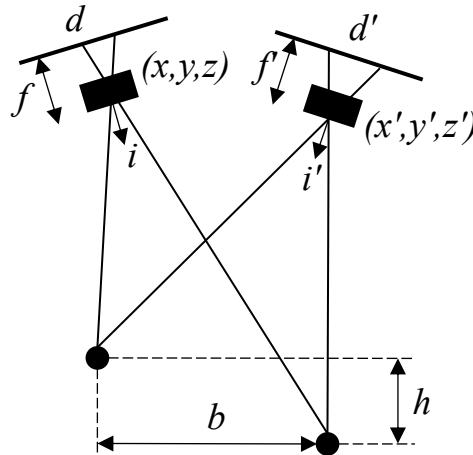


Рис. 1.11. Ілюстрація методу структура з руху (SfM)

SfM зазвичай поділяється на три типи: глобальний, інкрементний та гібридний.

Глобальний SfM [65] представлений на Рис. 1.12. Підхід оптимізує всі положення камери одночасно, використовуючи всі доступні переміщення. У глобальному SfM початкові положення камер оцінюються на основі принципів стерео зору та епіпольної геометрії. ВА [66] виконується лише один раз, що призводить до поліпшення ефективності системи. Процес глобального SfM включає, в основному, два етапи: оптимізацію положень камери та оптимізацію її орієнтацій. Точність оптимізації орієнтацій камери залежить від точності початкового розрахунку параметрів епіпольного геометричного графу.

Основні компоненти епіпольного геометричного графу:

- Центри камер (C_1 та C_2): Позиції двох камер у просторі.
- Точка в 3D просторі (P): Точка 3D сцени, яку спостерігають обидві камери.
- Точки зображення (p_1 та p_2): Проекції тривимірної точки P на площини зображень двох камер.
- Базисна лінія: Лінія, що з'єднує центри двох камер.

- Епіполярна площина: Площина, що проходить через тривимірну точку P та центри камер.
- Епіполі (e_1 та e_2): Точки, де базисна лінія перетинає площини зображень.
- Епіполярні лінії: Лінії на площинах зображень, вздовж яких лежать відповідні точки.

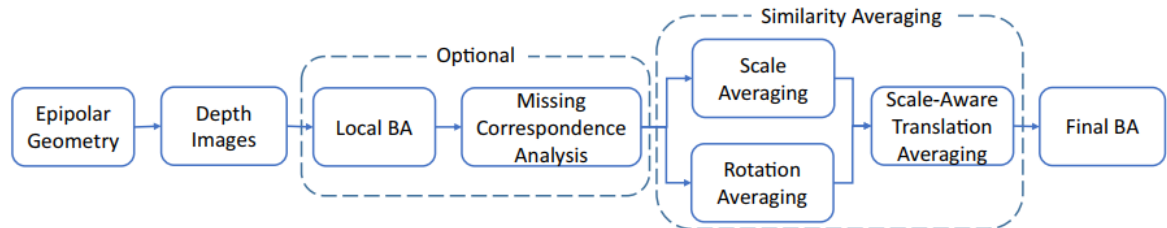


Рис. 1.12. Приклад алгоритму глобального SfM [65]

Інкрементний SfM [67] здатний обробляти великі обсяги даних. Цей підхід, зазвичай, використовуються саме для реконструкції розріджених об'ємів точок з неупорядкованого набору зображень. Метод включає кілька етапів, таких як калібрування камери, відстеження необхідних ознак, визначення положень камери, триангуляція та БА. Алгоритм можна описати наступним чином:

1. Ініціалізація вхідних зображень з урахуванням приблизної фокусної відстані.
2. Використання SIFT дескриптору [64] для пошуку необхідних ключових точок на зображенні.
3. Застосування порівняльного аналізу для визначення схожості знайдених ключових точок.
4. Визначення положення камери для першої пари зображень на основі п'яти точок.
5. Здійснення триангуляції для знаходження перших 3D точок, що формують об'єми точок у тривимірному просторі та виконання БА.
6. Використання методу прямого лінійного перетворення (Direct Linear Transform – DLT) для визначення положення кожної використаної камери.

7. Здійснення триангуляції для додавання нових точок до об'ємів точок з подальшим виконанням ВА для уточнення результатів.
8. Повторення кроків 5 та 6 до тих пір, поки не завершиться додавання нових камер. Для уточнення 3D об'єму точок модель оптимізується за допомогою глобального ВА (global bundle adjustment) [66].

Гібридний SfM використовує глобальний SfM для знаходження параметрів камери та інкрементний SfM для реконструкції розріджених об'ємів точок.

SfM підхід часто застосовують на початковому етапі 3D реконструкції. Отримані положення камери та розріджені об'єми точок подаються на вхід наступних методів, які уточнюють отримані дані для відновлення щільної та деталізованої тривимірної моделі середовища. Один з таких методів – багатовидове стерео зіставлення (Multi-View Stereo - MVS).

1.2.3. Багатовидове стерео співставлення

Алгоритми Multi-View Stereo (MVS) [68] приймають на вхід велику кількість зображень з даними про їх відповідні положення камери та повертають щільні тривимірні моделі з покращеною точністю. MVS підхід заснований на патчах (Patch-based Multi-view Stereo – PMVS) [69] дозволяє ефективно отримувати щільні об'єми точок. Метод включає три етапи: ініціалізацію, розширення та фільтрацію. Основна ідея PMVS – відтворення набору вирівняних патчів, що повністю покривають поверхню об'єкта. Модель патча є ключовим елементом PMVS. Процес реконструкції моделі за допомогою PMVS продемонстровано на Рис. 1.13. Патч P описується як тривимірний прямокутник, для якого конфігурація визначається центральними координатами, одиничним вектором нормалі, екземпляром зображення та колекцією зображень, на яких P розпізнається.



Рис. 1.13. Ілюстрація проміжних результатів MVS підходу заснованого на патчах

Ініціалізація патча передбачає наступні етапи:

1. Використання детектор Харріса та різницю гаусіанів (Difference of Gaussian – DoG), щоб виявити ключові точки на послідовності зображень.
2. Застосування техніки епіполярного порівняння для зіставлення ключових точок, що дають початкові відповідності.
3. Формування початкових патчів з початкових відповідностей за допомогою триангуляції.
4. Створення колекції розпізнаваних зображень для патча P за допомогою використання кутового тесту, що порівнює вектор спостереження направлений від камери до патчу та вектор нормалі відповідної поверхні.
5. Оновлення центру та нормалі шляхом зменшення оцінки фотометричної невідповідності.
6. Використання фотометричної диспаратності для оцінки ефективності створення патча P .

Розповсюдження патчу здійснюється поділом кожного зображення на систематичні сітки розміром $N \times N$ пікселів.

1. Збір всіх сусідніх клітинок зображення з кожного розпізнаного зображення в патчі P .
2. Створення нового патча P для зібраних клітинок зображення.

3. Оптимізація шляхом зменшення значення фотометричної диспаратності.
4. Оптимізація нового розпізнаного набору зображень за допомогою порівняння по глибині.
5. Визначення ефективності розширення нового патча.

Фільтрація патча виконується для оптимізації створених цільних об'ємів точок.

1.2.4. Реконструкція поверхні

Алгоритм відтворення поверхні за допомогою методу Пуассона (Poisson Surface Reconstruction – PSR) [70], використовується для створення цілісної структури. Цей метод розглядає завдання відтворення поверхні як просторову задачу Пуассона [70] і може бути описаний наступними етапами: дискретизація задачі, визначення векторного поля, вирішення рівняння Пуассона та відновлення ізоповерхні. Основні етапи відтворення поверхні за методом Пуассона включають:

1. Використання дерева октантів (octree) для ілюстрації ступеню функції розмірів функцій та дискретизації задачі вирішення структури Пуассона.
2. Для визначення векторного поля використовується функція вузла дерева октантів (octree), яка схожа на градієнтне поле функції індикатора.
3. Для вирішення рівняння Пуассона використовується описане векторне поле так, щоб градієнтне поле функції індикатора було суміжним з векторним полем.
4. Для того, щоб зобразити відповідну ізоповерхню використовується оцінена функція індикатора.

1.2.5. Підхід, який поєднує SfM, MVS та Реконструкцію поверхні

У [71] запропонований підхід 3D реконструкції, який поєднує в собі SfM, MVS та реконструкцію поверхні. Алгоритм умовно зображений на Рис. 1.14. Процедура 3D реконструкції дозволяє отримувати окремі 3D моделі, разом з розрідженим та щільним об'ємом 3D точок, а також грубі структури з колекції зображень. Цей метод можна використовувати в різних сферах та для виконання різноманітних завдань.

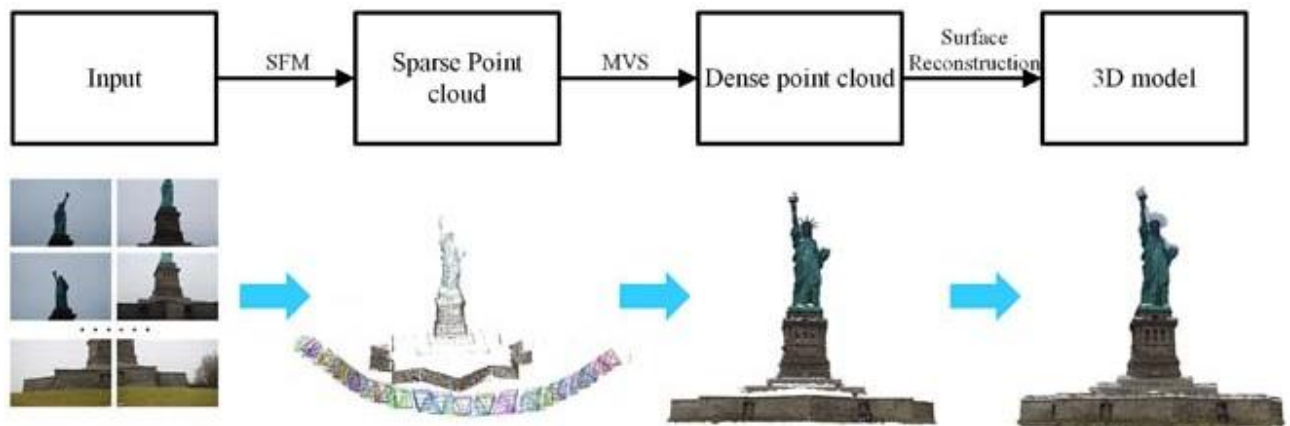


Рис. 1.14. Ілюстрація роботи методу, що поєднує SfM, MVS та реконструкцію поверхні [71]

Інкрементальний підхід SfM використовується для отримання розріджених об'ємів точок з вхідних зображень, при цьому обчислюються різні параметри для кожного зображення. Цей підхід включає етапи калібрування камери, відстеження особливостей, обчислення позицій камери, тріангуляції та ВА. Для отримання щільних об'ємів точок використовується підхід MVS на основі патчів, що включає ініціалізацію, розширення та фільтрацію патчів. Після отримання щільних об'ємів точок для створення моделі використовується підхід реконструкції поверхні за методом Пуассона, яка включає етапи дискретизації задачі, визначення векторного поля, розв'язання рівняння Пуассона та побудову ізоповерхні.

1.3. Сучасні методи реконструкції карт глибини та 3D реконструкції середовища

У сучасному світі нейронні методи для відновлення карт глибини та 3D реконструкції стали більш популярними порівняно зі стандартними та класичними підходами, такими як Structure from Motion (SfM) та Multi-View Stereo (MVS). Причини цього включають кілька ключових аспектів:

- *Висока точність і деталізація.* Нейронні мережі, зокрема глибокі нейронні мережі, можуть навчатися на великих наборах даних і знаходити складні, нелінійні залежності між зображеннями та їх тривимірними представленнями. Це дозволяє їм досягати більш високої точності та деталізації в порівнянні з традиційними методами, які часто базуються на простих геометричних припущеннях. Наприклад, на відміну від класичних підходів, сучасні методи дозволяють отримувати якісні карти глибин та 3D реконструкцію навіть в областях з однорідними та періодичними текстурами.
- *Автоматизація і узагальнення.* Глибокі нейронні мережі здатні автоматично витягати особливості та патерни з вхідних даних, що зменшує потребу в ручному налаштуванні параметрів та попередньому обробленні даних, як це часто потрібно у традиційних методах. Це робить їх більш універсальними та здатними до узагальнення на нові сцени та умови зйомки.
- *Зменшення чутливості до шумів.* Нейронні методи мають вбудовану здатність до зменшення впливу шумів та артефактів в даних завдяки використанню навчання на великих і різноманітних наборах даних. Це дозволяє отримувати більш стабільні результати навіть при наявності шумів у вхідних зображеннях.
- *Інтеграція різних джерел даних.* Нейронні мережі можуть інтегрувати інформацію з різних даних, таких як RGB-зображення, карти глибини, та

інші сенсорні дані, для створення більш комплексних і точних 3D моделей.

1.3.1. HighRes-MVSNet

В [72] використано архітектуру моделі глибокого навчання для 3D реконструкції із зображень високої роздільної здатності. Традиційні MVS-техніки для уточнення значення глибини кожного пікселя використовують обчислені ознаки та зв'язки між кількома видами сцени [73]. Використання ознак, виділених нейронною мережею під час навчання, є достойною альтернативою. Архітектура HighRes-MVSNet зосереджується на зниженні вимог до пам'яті необхідної для використання великої кількості даних, які доступні завдяки сучасним камерами з високою роздільною здатністю зображень. Підхід використовує архітектуру енкодера-декодера, зображену на Рис. 1.15. В енкодері система спочатку виконує 3 згортки у початковому шарі, після чого йде шар об'єднання регіонів даних (pooling) і ще один згортковий шар. Таким чином, розмір обсягу ознак зменшується до однієї восьмої від вхідних даних зображення.

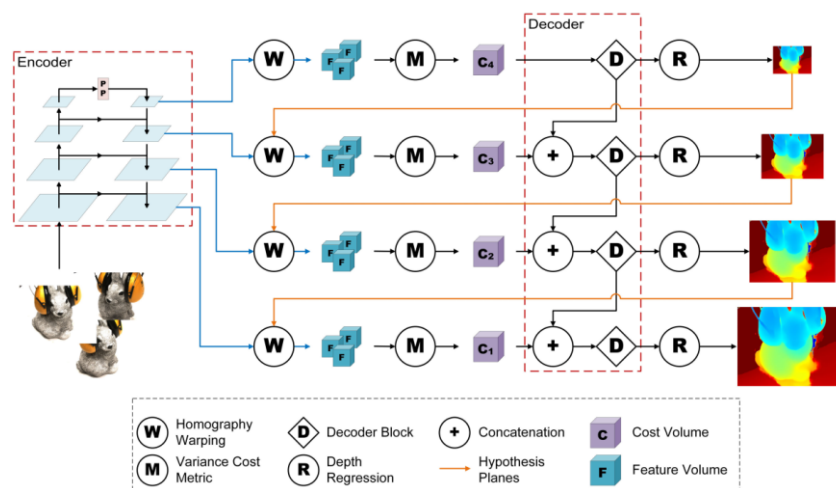


Рис. 1.15. Архітектура мережі HighRes-MVSNet [72]

Потім використовується архітектура U-Net для вилучення ознак на декількох масштабах, і на кожному масштабі ознаки проєктуються в нижчий вимірний підпростір для контролю розміру вихідної карти ознак. Диференційована функція гомографічного викривлення використовується для

побудови 3D об'ємів цільових значень з попередньо вилучених ознак на кожному масштабі. Функція викривлення задається наступним чином:

$$H_i(d) = K_i \cdot R_i \cdot \left(I - \frac{t_0 - t_i \cdot n_0^T}{d} \right) \cdot R_0^T \cdot K_0^T, \quad (1.4)$$

де $H_i(d)$ є гомографією між i -тою картою ознак і картою ознак на глибині d . K_i , R_i та t_i відповідають параметрам камери, що вказують на поточний кадр, n_0 – головна вісь поточної камери. N об'ємів ознак F_i збираються у об'єм цільових значень C за допомогою метрики втрат на основі варіації:

$$C = \frac{\sum_{i=1}^N (F_i - \bar{F}_i)^2}{N}, \quad (1.5)$$

де \bar{F}_i – усереднений об'єм ознак.

Об'єми цільових значень декодуються від грубого до детального рівня та об'єднують усі виходи з найдетальнішим масштабом. Декодерна система має чотири блоки, кожен з яких відповідає за результат на одному з чотирьох етапів. Кожен блок декодера складається з шести 3D-згорткових блоків, кожен з яких містить дві 3D-згортки з залишковим з'єднанням. Декодер надає два виходи: об'єм цільових значень, який об'єднується з вхідними даними наступного етапу, та класифікований об'єм цільових значень, який обробляється шаром softmax і регресією глибини для створення карти глибини, що ініціює об'єми ознак наступного етапу.

Класифікація здійснюється за допомогою шару 3D-згортки, за яким іде шар ReLU та ще один шар 3D-згортки. На найбільш грубому етапі результат створюється лише з сирого об'єму цільових значень. Використовуючи концепцію каскадного об'єму цільових значень, ці об'єми цільових значень формуються в більш точні діапазони глибин залежно від попередньої оцінки. Глибина обчислюється на кожному масштабі шляхом збільшення роздільної здатності класифікованих об'ємів цільових значень до необхідного розміру виходу перед використанням регресії глибини.

1.3.2. 3D-FHNet

Тривимірний ієрархічний мережний об'єднання був представлений в [74]. В її основі лежать ідеї використання як техніки комбінування ознак з кількох видів, так і стратегія ієрархічного прогнозування. Це забезпечує можливість об'єднання реконструкції з однієї та декількох точок спостереження з метою отримання точних результатів. Техніка комбінування ознак спрямована на постійне підвищення якості реконструкції моделі зі збільшенням кількості нових ракурсів. Ієрархічна стратегія прогнозування впроваджена в мережу для точного відтворення дрібних деталей об'єктів. Модель представлена у вигляді воксельного подання, де кожен воксель позначається нулем або одиницею для кожної воксельної сітки. На Рис. 1.16 показана архітектура системи моделі.

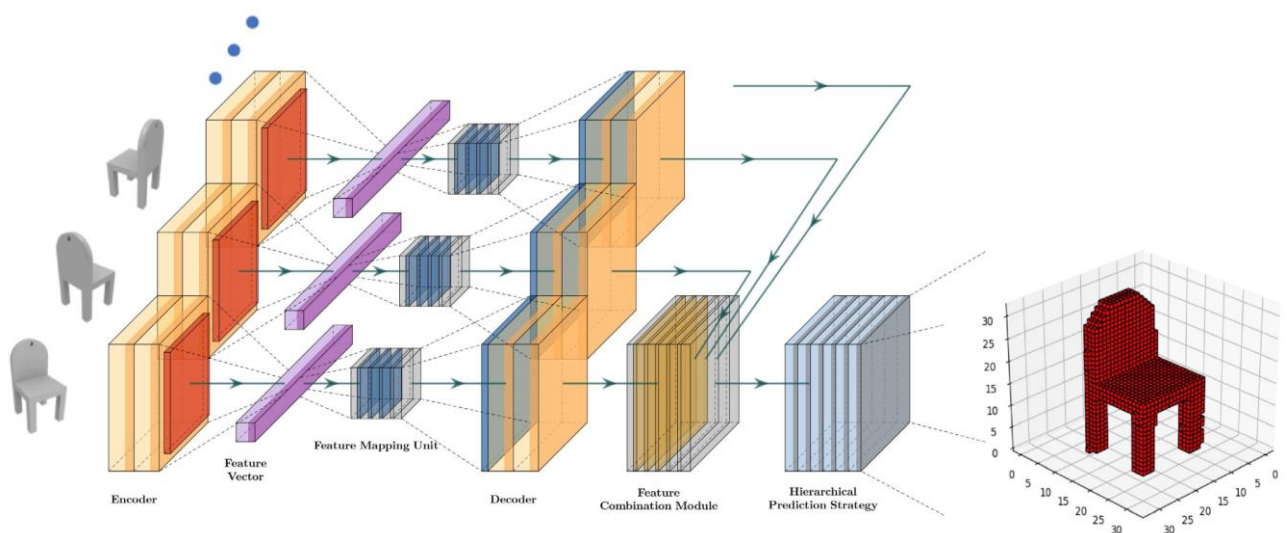


Рис. 1.16. Архітектура моделі 3D-FHNet [74]

Вхідними даними моделі можуть бути як одне, так і кілька зображень на яких присутній об'єкт представлений з різних точок спостереження. На етапі навчання моделі можна подавати численні зображення одного й того самого об'єкта або сцени. Ці зображення проходять через мережу енкодера, яка вилучає ознаки з поданих даних. Мережа використовується для вилучення ознак із зображень з роздільною здатністю 128×128 пікселів. Підхід використовує залишковий двовимірний енкодер. Кожне вхідне зображення проходить через шість залишкових згорткових блоків енкодера. Кожен із цих блоків виконує три згортки та одну операцію об'єднання регіонів даних (pooling). Для кожного блоку

дані проходять одночасно двома шляхами. Один шлях обробляє дві згортки, а другий шлях – згортку розміром 1×1 . Ці процеси супроводжуються функцією активації ReLU. Ознаки, отримані з обох шляхів, об'єднуються та проходять через шар max-pooling. Нарешті, дані перетворюються у вектор ознак.

Модуль відображення ознак використовується для перетворення двовимірних ознак у тривимірні, відображаючи вектори ознак, отримані з енкодера. Кожен вектор ознак проходить через повнозв'язний шар, а потім через модуль відображення ознак.

Тривимірна інформація декодується за допомогою залишкового тривимірного декодера. Декодер перетворює тривимірні ознаки в тривимірні об'єми. Він приймає ознаки з модуля відображення ознак та пропускає їх через шість тривимірних залишкових декодерних блоків. Результат потім нормалізується шаром softmax до тривимірного об'єму передбаченої ймовірності. Кожен блок містить три операції зворотної згортки та одну операцію розпакування (unpooling). Дані одночасно проходять двома шляхами. Один із шляхів містить дві тривимірні згортки, а інший шлях – згортку $1 \times 1 \times 1$. Усі ці процеси супроводжуються функцією активації ReLU. Дані надходять до наступного операційного блоку після об'єднання та проходження через шар розпакування (unpooling).

Для об'єднання ознак усіх отриманих зображень використовується модуль інтеграції ознак. Людина може отримати уявлення про об'єкт, рухаючись навколо нього та спостерігаючи його з різних ракурсів. Аналогічно, модель може отримати поточну передбачену ймовірність зайнятості вокселів, розглядаючи зображення. Коли кількість вхідних видів об'єкту збільшується, кількість воксельних сіток також зростає, відповідно точність моделі покращується.

Модель використовує ієрархічну стратегію прогнозування для виведення зайнятості вокселів у виді 0 – 1. Значення 0 вказує на відсутність зайнятості, а значення 1 вказує на зайнятість відповідного вокселю. Спочатку модель встановлює поріг, і воксельна сітка з передбаченою ймовірністю, рівною або більшою за поріг, буде класифікована як зайнята, а ті, що мають менше значення,

як незайняті. Це також допомагає відновити дрібні деталі об'єкта. Таку метрику якості, як точність, можна обчислити, порівнюючи передбачену зайнятість вокселів 0 – 1 з фактичною зайнятістю вокселів.

1.3.3. ATLAS

Одним з варіантів представлення 3D об'єкта, або площини в комп'ютерній графіці є усічена знакова функція відстані (Truncated Signed Distance Function – TSDF). Метод 3D-реконструкції сцени, що безпосередньо регресує TSDF із набору RGB-зображень та відповідних положень камери представлений у [14].

ATLAS приймає на вхід послідовність RGB-зображень довільної довжини, внутрішні параметри камери та положення камери відповідні до кожного кадру. Зображення проходять через основу 2D CNN для вилучення ознак. Потім ці ознаки зворотно проєктуються в 3D-воксельний об'єм і акумулюються та усереднюються з вже існуючими даними. Після того як ознаки зображень об'єдналися в 3D, відбувається безпосередня регресія TSDF за допомогою 3D CNN (Рис. 1.17).

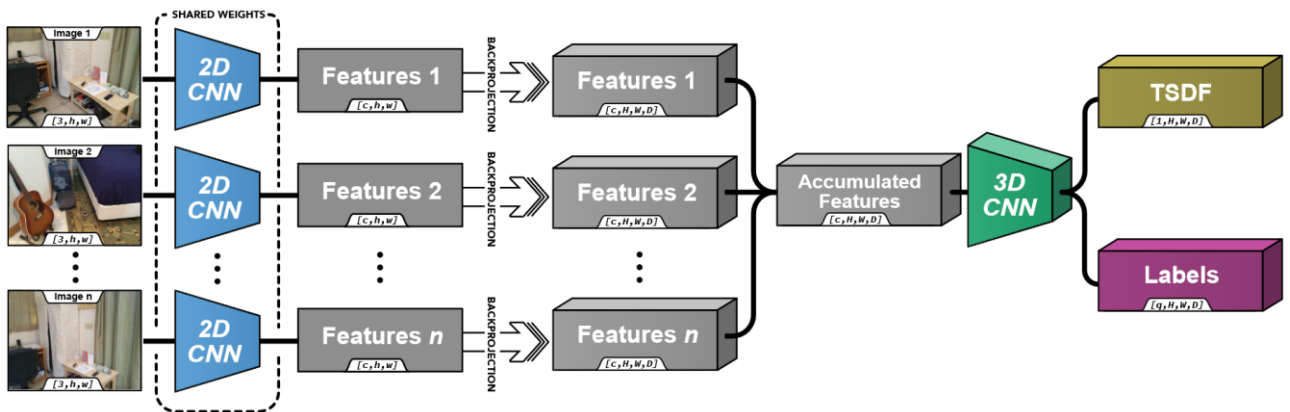


Рис. 1.17. Архітектура моделі ATLAS [14]

Після акумуляції ознак у воксельному об'ємі використовується тривимірна згорткова мережу енкодера-декодера для уточнення ознак і регресії вихідного TSDF. Кожен шар енкодера і декодера використовує набір залишкових блоків розміром $3 \times 3 \times 3$. Зменшення масштабу реалізується за допомогою згортки $3 \times 3 \times 3$ з кроком 2, а збільшення масштабу використовує трилінійну інтерполяцію з наступним застосуванням згортки $1 \times 1 \times 1$ для зміни розміру ознак. Розмір ознак

подвоюється при кожному зменшенні масштабу і зменшується вдвічі при кожному збільшенні масштабу. Усі згорткові шари супроводжуються batchnorm та активацією ReLU.

На верхньому шарі енкодера-декодера використовується згортка $1 \times 1 \times 1$, що супроводжується функцією активації tanh для регресії кінцевих значень TSDF.

Крім того, модель має проміжні виходи на кожній декодованій роздільній здатності перед збільшенням масштабу. Ці додаткові виходи використовуються як для проміжного контролю, щоб допомогти мережі швидше навчатися, так і для покращення якості наступних ітерацій.

1.3.4. SimpleRecon

Багато сучасних методів 3D реконструкції для підвищення якості застосовують важкі тривимірні згорткові шари, тим саме обмежуючи своє застосування в умовах обмежених обчислювальних ресурсів. На відміну від них SimpleRecon [15] робить акцент на можливість застосування в умовах обмежених обчислювальних ресурсів, при цьому зберігає високу якість прогнозування глибини. Традиційний та доступний підхід об'єднання карт глибини у поєднанні із запропонованим якісним відновленням глибини з кількох ракурсів призводить до точних 3D реконструкцій.

На вхід методу подаються: опорне зображення I_0 , набір вхідних зображень I_n , де $n \in \{1, \dots, N-1\}$, а також внутрішні параметри камер та відносні позиції камер. Під час навчання використовуються карти глибини D_{gt} , що відповідають вхідним RGB-зображенням. Під час тестування модель виконує передбачення щільних карт глибини \hat{D} для кожного опорного зображення.

SimpleRecon доповнює архітектуру енкодера-декодера для прогнозування глибини об'ємом цільових значень (Рис. 1.18).

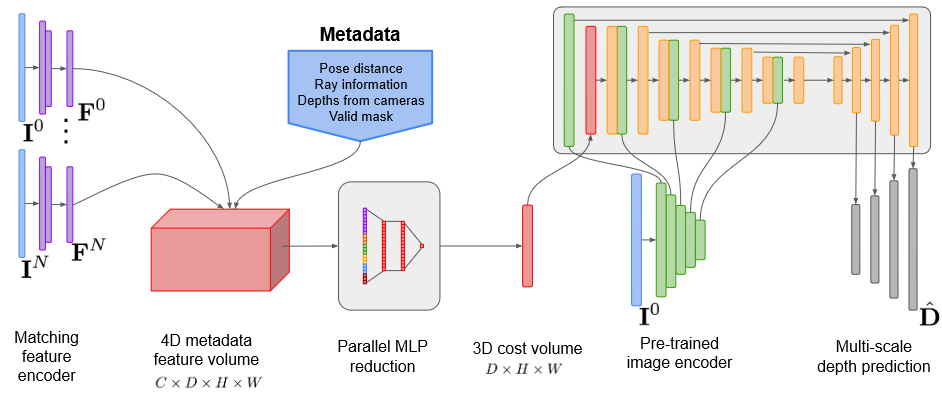


Рис. 1.18. Архітектура моделі SimpleRecon [15]

Енкодер зображень вилучає відповідні ознаки з опорних зображень, що подаються у об'єм цільових значень. Вихід об'єму цільових значень обробляється за допомогою двовимірної згорткової мережі енкодера-декодера, доповненої ознаками вилученими за допомогою окремого попередньо навченого енкодера зображень. Ключова ідея підходу полягає у включенні доступних метаданих до об'єму цільових значень разом із типовими глибинними ознаками зображення, що дозволяє мережі використовувати корисну інформацію, таку як геометричні дані та відносна позиція камер. Ці метадані дозволяють мережі краще визначати відносну важливість кожного вхідного зображення для оцінки глибини для певного пікселя.

Об'єм цільових значень є чотиривимірним тензором розмірністю $C \times D \times H \times W$, де для кожного просторового положення (k, i, j) існує вектор ознак розмірністю C , де k — індекс площини глибини. Вектор ознак C складається з ознак опорного зображення та набору ознак вхідних зображень, що деформуються враховуючи специфічні метадані.

Мережа базується на двовимірній згортковій архітектурі енкодера-декодера. Об'єднання об'єму цільових значень відбувається за рахунок підсумовування результатів добутку точок між опорним зображенням і кожним вхідним зображенням та дає результати, які конкурентоспроможні з сучасними методами оцінки глибини.

В якості енкодера зображень та енкодера відповідності ознак використовується невеликий, але потужніший енкодер EfficientNetv2 S [75]. Для

ефективного створення карт відповідності ознак застосовуються перші два блоки ResNet18 [76].

В SimpleRecon об'єднання ознак зображення в енкодер об'єму цільових значень відбувається на зразок DeepVideoMVS [77]. Глибинні ознаки зображення об'єднуються на кількох масштабах, додаючи пропускні з'єднання між енкодером зображення та енкодером об'єму цільових значень на всіх роздільних здатностях.

Функція втрат представлена у вигляді:

$$L = L_{depth} + \alpha_{grad} L_{grad} + \alpha_{normals} L_{normals} + \alpha_{mv} L_{mv}, \quad (1.6)$$

де: L_{depth} – функція втрат регресії глибини; L_{grad} – функція втрат багатомасштабного градієнту та нормалей; L_{mv} – функція втрат багатовидової регресії глибини; $\alpha_{grad} = \alpha_{normals} = 1.0$ та $\alpha_{mv} = 0.2$ – коефіцієнти підібрані експериментально на валідаційному наборі даних.

1.3.5. Marigold

Підхід Marigold [78], базується на використанні дифузійної моделі та пропонує протокол тонкого налаштування для монокулярної реконструкції карти глибини. Враховуючи щільну карту глибини, що повертає алгоритм, та її високу точність, підхід також застосовується і для задач 3D реконструкції (Рис. 1.19).

Основний принцип підходу полягає у використанні багатих візуальних знань, що зберігаються в сучасних генеративних візуальних моделях. Marigold, розроблена на основі попередньо навченої Stable Diffusion моделі та доналаштована з використанням синтетичних даних.

Однією з основних цілей Marigold є ефективність навчання, оскільки дифузійні моделі зазвичай потребують багато ресурсів для навчання. Тому запропонована модель базується на попередньо навченій text-to-image LDM (Stable Diffusion v2 [79]), яка навчилася дуже хорошим апостеріорним ознакам зображень на наборі даних LAION-5B [80].

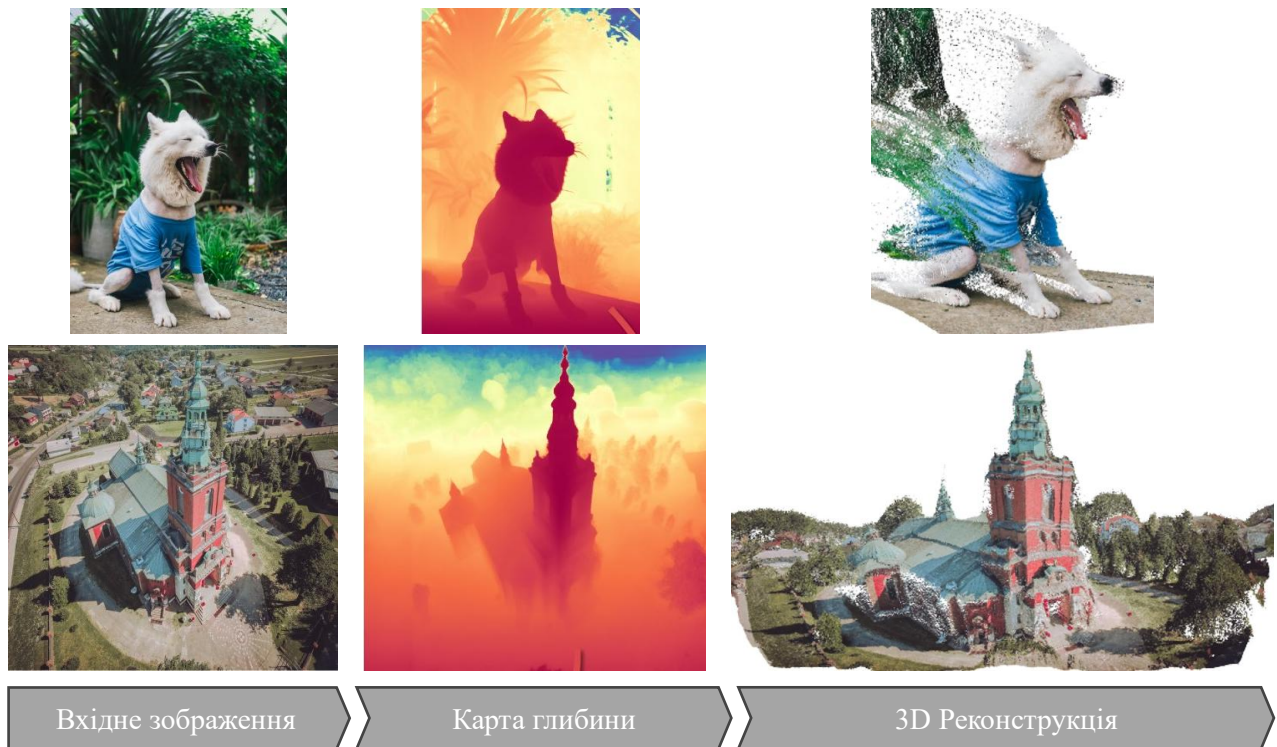


Рис. 1.19. Приклади 3D реконструкції із застосуванням Marigold [78]

Для кодування зображення x та його відповідної карти глибини d в латентний простір, з метою навчання умовного розшумлювача, використовується оригінальний зафіксований варіаційний автоенкодер (VAE) Stable Diffusion. Карта глибини дублюється в три канали, щоб імітувати RGB-зображення та подається на енкодер. Доналаштовується лише U-Net, оптимізуючи стандартну дифузійну задачу відносно латентного коду глибини. Відповідність співставлення карти глибини до зображення досягається шляхом об'єднання двох латентних кодів перед їх подачею в U-Net. Перший шар U-Net модифіковано для прийняття об'єднаних латентних кодів. Підхід доналаштування Marigold представлений на Рис. 1.20.

Вхідне зображення x кодується за допомогою оригінального варіаційного автоенкодера (VAE) Stable Diffusion у латентний код $z^{(x)}$ і об'єднується з латентним кодом глибини $z_t^{(d)}$ перед подачею в модифіковану доналаштовану U-Net мережу на кожній ітерації розшумлення. Після T ітерацій розшумлення, отриманий латентний код глибини $z_0^{(d)}$ декодується в зображення, три канали якого усереднюються для отримання кінцевої оцінки \hat{d} .

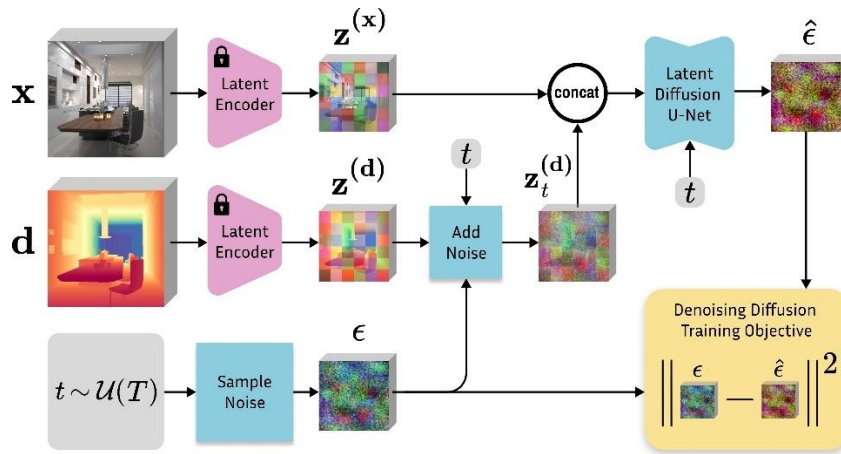


Рис. 1.20. Огляд підходу доналаштування Marigold [78]

Загальна схема виконання підходу представлена на Рис. 1.21.

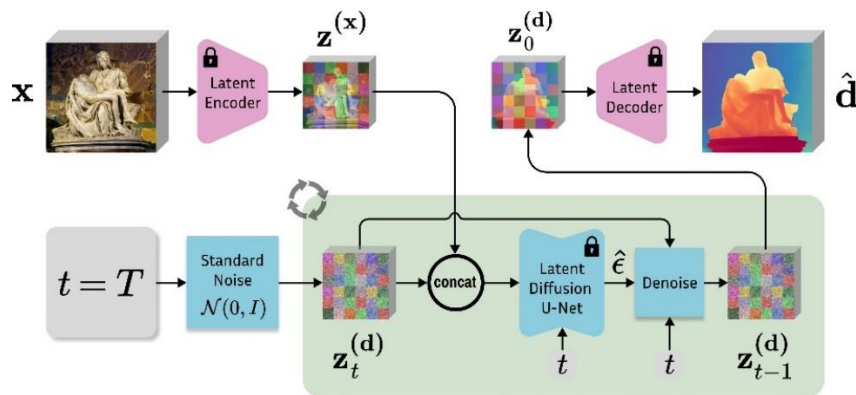


Рис. 1.21. Загальна схема Marigold для реконструкції карти глибини [78]

Основний недолік Marigold – тривалий час виконання, що робить цей підхід непридатним для систем, які оперують в умовах реального часу.

1.4. Методи реконструкції та рендерингу

Розглянемо деякі новітні методи орієнтовані на фотореалістичну 3D реконструкцію середовища та синтез новітніх видів сцени з метою розширення наявних наборів даних.

1.4.1. Neural Radiance Fields (NeRF)

Останні дослідження показали, що NeRF здатний успішно синтезувати нові види навіть для складних сцен [81]. Представлення сцена забезпечується повнозв'язною глибокою нейронною мережею (MLP). На вхід подається

п'ятивимірною векторною функцією з аргументами: \vec{x} – просторовим положенням (x, y, z) та \vec{d} – напрямом спостереження (θ, ϕ) . На виході мережа повертає щільність об'єму σ та RGB-колір \vec{c} для кожного пікселя. Формально модель задають так [81]:

$$F_w : (\vec{x}, \vec{d}) \rightarrow (\vec{c}, \sigma) \quad (1.7)$$

Для навчання NeRF потрібен набір RGB-зображень однієї сцени, знятих під різними кутами спостереження, а також положення камери та її внутрішні параметри. У процесі навчання рендерять відповідні види сцени та мінімізують фотометричну похибку між синтезованими та зображеннями, що спостерігались.

Схему рендерингу показано на Рис. 1.22. Спершу через кожен піксель проводять промінь камери та відбирають низку точок уздовж цього променя. Далі ці точки подають до багатошарового перцептрона, який передбачає для них колір і щільність. На завершальному етапі застосовують класичну об'ємну візуалізацію [82], агрегуючи внески кольорів та щільностей усіх вибірок для отримання підсумкового значення кожного пікселя.

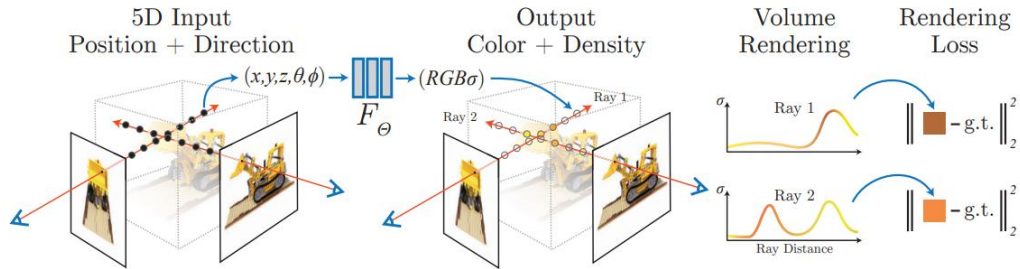


Рис. 1.22. Процес рендерингу за допомогою NeRF [81]

Під час навчання NeRF мінімізується фотометрична функція втрат. Типово використовують дві мережі: «грубу» та «точну» [81], але для простоти розглянемо лише «грубу» підмережу. Нехай маємо M зображень (I_1, \dots, I_M) . Тоді метою навчання є оптимізація такої цільової функції синтезу:

$$L_F = \sum_{i=1}^M \sum_u \|\hat{I}_i(u; w) - I_i(u)\|_2^2, \quad (1.8)$$

де w – параметри моделі, що залежать від напрямів спостереження, u – координати пікселів, $\hat{I}_i(u; w)$ – синтезоване RGB-значення кольору у пікселі u .

Для отримання значень кольорів пікселів, через які проходять промені, застосовують класичні техніки рендерінгу [82]. Формула для обчислення кольору вздовж променя $r(t) = o + td$ має наступний вигляд:

$$C(r) = \int_{t_n}^{t_f} T(t) \sigma(r(t)) c(r(t), d) dt, \quad (1.9)$$

де t_n та t_f значення діапазону, в якому змінюється t , $r(t)$ – вхідний промінь, $\sigma(r(t))$ – об’ємна щільність, яку також можна інтерпретувати як імовірність закінчення променя в точці t , $c(r(t), d)$ – RGB колір променя в точці t , $T(t)$ – коефіцієнт, що характеризує проникність променю до точки t тобто імовірність того, що промінь пройде від точки t_n до t , не стикаючись з іншими частинками:

$$T(t) = \exp\left(-\int_{t_n}^t \sigma(r(s)) ds\right) \quad (1.10)$$

Для обчислення глибини в пікселі $D(r)$ використовують вихідний параметр (1.7) – щільність σ для розрахунку очікуваної відстані завершення вздовж променя:

$$D(r) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) t_i, \quad (1.11)$$

де t_i – відстань від камери вздовж променя, δ_i – довжина i -го відрізка вздовж променя між двома сусідніми вибірками.

1.4.2. 3D Gaussian Splatting (3DGS)

Neural Radiance Fields підхід [81] здійснив перелом у задачі синтезу нових видів сцени, забезпечивши фотореалістичний рендеринг за набором вхідних зображень. Водночас обчислювальні витрати NeRF обмежують їх практичне застосування в режимах реального часу. У статті [83] запропоновано підхід 3D Gaussian Splatting (3DGS), який досягає швидкостей рендерингу в реальному часі, зберігаючи при цьому прийнятну візуальну якість.

3DGS поєднує найкращі властивості нейронних і точкових (point-based) методів рендерингу. Метод схематично представлено на Рис. 1.23. 3DGS використовує розріджений об’єм точок, отриманий за допомогою SfM підходу для ініціалізації 3D-гаусіанів з метою представлення сцени. Отримані гаусіани

далі ітеративно проєктуються у вид камери, до них застосовується диференційований растеризатор і результат оптимізується через адаптивний контроль щільності.

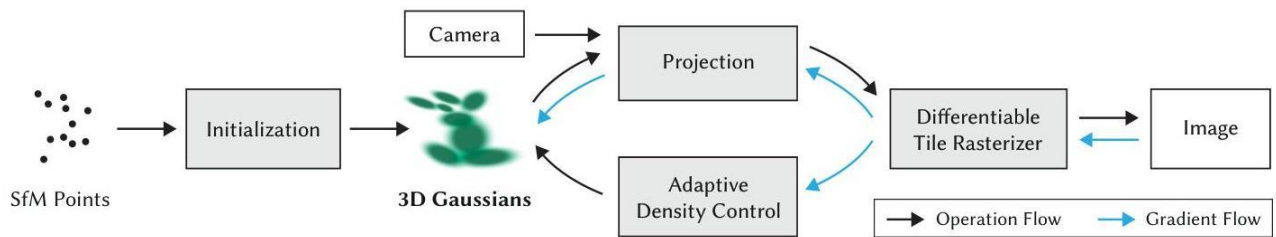


Рис. 1.23. Процес рендерингу за допомогою 3DGS [83]

На відміну від традиційних точкових підходів, що потребують щільного об'єму точок отриманого за допомогою MVS, 3DGS потребує лише розрідженого об'єму точок, отриманих з SfM, які зазвичай доступні після калібрування камери. Кожен 3D-гаусіан визначається:

- 3D-позицією;
- 3D-матрицею коваріації, що описує анізотропну форму;
- непрозорістю (opacity);
- коефіцієнтами сферичних гармонік для виглядозалежного представлення.

Анізотропність гаусіанів дозволяє краще моделювати складні геометрії, як-от тонкі структури. Матриця коваріації кожного гаусіана задається як:

$$\Sigma = RSS^T R^T, \quad (1.12)$$

де R – матриця повороту, а S – матриця масштабування. Таке представлення дозволяє розтягувати й орієнтувати гаусіани в 3D-просторі для ефективного наближення поверхонь.

Процес оптимізації ітеративно уточнює параметри кожного гаусіана, зменшуючи розбіжність між зрендереними та еталонними зображеннями. Ключовою інновацією [83] є механізм адаптивного контролю щільності, який динамічно додає або вилучає гаусіани залежно від якості рендерингу. Коли область недостатньо реконструйована (недостатня щільність гаусіанів), гаусіани клонуються для підвищення деталізації. Для занадто реконструйованих регіонів

(надлишкова щільність гаусіанів), великі гаусіани розщеплюються на менші. Такий адаптивний підхід продемонстровано на Рис. 1.24. Він гарантує фокусування обчислювальних ресурсів на ділянках, які потребують більшої деталізації.

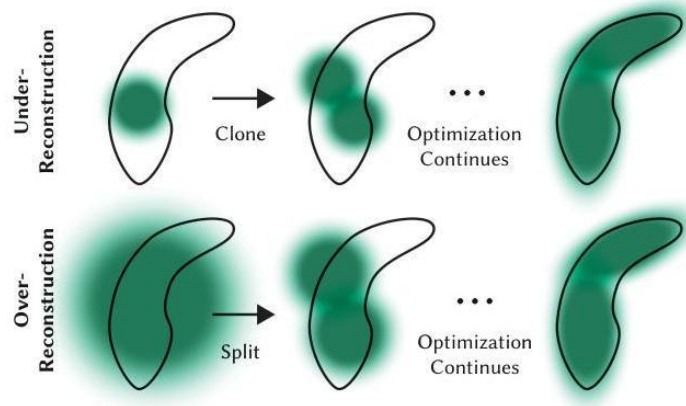


Рис. 1.24. Механізм адаптивного контролю щільності [83]

Алгоритм рендерингу є критичним компонентом системи, що забезпечує і швидке навчання, і рендеринг у реальному часі. Процес включає:

- проєціювання 3D-гаусіанів у 2D-екранний простір;
- сортування гаусіанів за глибиною;
- застосування плиткового диференційованого растеризатора;
- ефективний зворотний прохід (backward pass) для оптимізації.

Для кожного пікселя колір обчислюється як

$$C = \sum_{i \in N} c_i a_i \prod_{j=1}^{i-1} (1 - a_j), \quad (1.13)$$

де c_i – колір i -го гаусіана, a_i – його непрозорість, а N – множина гаусіанів, що впливають на даний піксель.

Диференційовність усього підходу дає змогу виконувати наскрізну (end-to-end) оптимізацію. Алгоритм рендерингу забезпечує високу швидкість понад 100 кадрів на секунду за роздільної здатності 1080p, що робить його придатним для застосувань у реальному часі.

Порівняння NeRF та 3DGS підходів представлено в Табл. 1.2.

Табл. 1.2. Порівняльний аналіз NeRF та 3DGS підходів

Характеристика	3DGS	NeRF
Представлення	Явне – набір мільйонів дискретних, навчених 3D-гаусіанів	Неявне – об’ємна функція, представлена нейронною мережею
Рендеринг	Використовує швидкий, прискорений на GPU процес растеризації для «сплатингу» гаусіанів на 2D-площину зображення	Використовує обчислювально витратний процес «ray marching», коли нейронна мережа опитує тисячі точок уздовж променя
Швидкість навчання	Значно швидше: час навчання часто вимірюється хвилинами	Потребує тривалого навчання – для збіжності можуть знадобитися години
Швидкість виконання	Досягає рендерингу в реальному часі на високих роздільностях (наприклад, понад 100 FPS)	Значно повільніше: часто потрібні секунди або хвилини, щоб зрендерити одне зображення
Використання пам’яті	Може бути вимогливим до пам’яті через зберігання явних гаусіанових примітивів	Ощадливіший щодо пам’яті під час виконання, але потребує значних ресурсів на навчання
Якість синтезу новітніх видів сцени	Забезпечує високу візуальну різкість і детальність, особливо на типових поверхнях. Втім, якість залежить від щільності та розміщення дискретних гаусіанів, що може спричинити «плаваючі» артефакти, зокрема в складних або динамічних сценах. Геометричне представлення інколи може бути менш точним	Завдяки неперервній об’ємній функції, представленій нейронною мережею, перевершує здатністю створювати різкі нові види сцени, що дає чистіші й точніші реконструкції. Більш стабільний і краще відновлює геометрію за обмеженої кількості вхідних зображень

Оскільки NeRF підходи, не зважаючи на свою обчислювальну складність, забезпечують кращу якість реконструкції, особливо для складних сцен, то далі будемо розглядати саме їх та їх похідні підходи для задачі 3D реконструкції та синтезу новітніх видів сцени, з метою розширення наявних наборів даних.

1.4.3. NeRF похідні методи

За наявності набору знімків сцени, зроблених із різних положень камери, NeRF відновлює її тривимірну геометрію та дає змогу синтезувати нові ракурси. Водночас це не єдиний шлях розв’язання задачі генерування нових видів сцени. Застосовують також оптимізацію подання сцени на тривимірній сітці [84] і підходи на основі глибинних мереж, які відображають координати XYZ у функції знакової відстані (sign distance functions, SDF) [85, 86]. Однак висока обчислювальна складність таких методів зумовлює залежність від якісних еталонних 3D-даних і слабку масштабованість під час синтезу зображень високої роздільної здатності. На відміну від них, NeRF оптимізує представлення сцени у формі неперервної диференційованої функції, що дозволяє навчати модель наскрізним шляхом (end-to-end).

Попри переваги, базовий NeRF [81] має істотні обмеження через тривалий час оптимізації та жорсткі вимоги до точності положень камер. У [87] автори запропонували модель BARF, що дає можливість пом’якшити вимоги до точності положень камери шляхом їх додавання безпосередньо до процесу оптимізації. Водночас і NeRF, і BARF залишаються прив’язаними до одиничної сцени. Вони не узагальнюються на інші середовища і здатні синтезувати зображення лише для тієї сцени, на якій були навчені.

Частково цю проблему адресують варіанти NeRF із залученням multi-view stereo (MVS), зокрема MVSNeRF [88] та NerfingMVS [89]. У MVSNeRF [88] модель можна попередньо навчити на одному датасеті, а далі донавчити на іншому. Таким чином, на відміну від інших робіт, пов’язаних із синтезом зображень за допомогою NeRF, архітектура MVS полегшує пошук міжкадрових відповідностей, а поєднання з 3D-згортковими мережами сприяє кращому узагальненню на нові набори даних. Головна мета NerfingMVS [89] – підвищити якість карт глибин, що водночас покращує здатність NeRF генерувати нові види сцени. Автори спершу оптимізують монокулярну мережу для оцінювання глибини на цільовій сцені, донавчаючи її на розріджених картах відстаней, отриманих методом structure-from-motion (SfM) за допомогою COLMAP [90].

Вибір точок уздовж променів у процесі оптимізації NeRF керується картою помилок синтезованих зображень, завдяки чому результуючі карти глибини стають точнішими й чіткішими.

Ще один напрям, де використовують COLMAP [90] – DDPNeRF [91]. Автори використовують побудову щільних карт глибини та карт невизначеності, які надалі слугують для регуляризації та оптимізації NeRF. Цей метод дозволяє зменшити потребу генерації нових видів сцени у вхідних даних до 18-36 зображень. DDPNeRF близький до NerfingMVS [89], однак додатково виконується заповнення карт відстаней, отриманих із COLMAP, та явне врахування невизначеності.

Усі розглянуті вище методи реконструкції 3D-середовища й синтезу зображень передбачають статичність сцени, а отже не моделюють змінне освітлення – це їхній принциповий недолік. Загалом NeRF-похідні моделі працюють за припущенням статичної геометрії сцени, хоча на практиці умови часто динамічні. Neural Scene Flow Fields (NSFF) [92] пропонує варіант NeRF для динамічних сцен. Автори моделюють динамічну сцену, модифікувавши представлення сцени так, щоб одночасно описувати середовище, геометрію та рух як неперервні часові залежні функції, що дозволяє інтерполяцію змін як у просторі, так і в часі. На відміну від [92], підхід запропонований в цьому розділі розглядає часову залежність NeRF для статичних сцен із динамічним освітленням [22].

Розглядаючи NeRF як джерело доповнення даних для подальшого навчання моделей розуміння сцени (наприклад, для передбачення глибини), варто враховувати й альтернативу на базі фотореалістичного рендерингу. У [93] представлено синтетичний набір Hypersim для задач розуміння сцени та методику його отримання. Завдяки фізично коректному моделюванню освітлення й розсіювання світла ключовою перевагою даного підходу є висока фотореалістичність зображень. Недолік полягає у вимозі до наявності якісних фотореалістичних 3D-моделей сцен та об'єктів, що обмежені у відкритому доступі. У цьому контексті перевага NeRF – можливість реконструкції та

генерації даних із відносно невеликої кількості лабораторно отриманих зображень.

1.5. Сучасні підходи реконструкції карт глибини та 3D реконструкції середовища орієнтовані на кінцеві пристрої

Розглянемо деякі новітні методи орієнтовані на реконструкцію карт глибини та 3D реконструкцію середовища, що ставлять за мету оптимізацію під кінцевий пристрій користувача.

1.5.1. DispNet

У роботі [94], окрім набору даних для задача прогнозування диспаратності, оптичного потоку та потоку сцени, було запропонували архітектуру моделі DispNet. Дослідники адаптували успішну кодувальню-декодувальню архітектуру FlowNet [95] для задачі прогнозування диспаратності сцени. Ця модель залишається актуальною і по сьогоднішній день для вирішення задачі прогнозування глибини та 3D реконструкції сцени. В [94] запропоновано кілька варіантів мереж:

- **DispNet:** згорткова мережа для оцінювання диспаратності, яка в декодері додає додаткові згорткові шари між операціями деконволюції. Така модифікація формує гладкіші та краще регуляризовані карти диспаратності порівняно з оригінальною архітектурою FlowNet. Архітектуру DispNet можна описати наступним чином [94]: кодувальна (encoder) частина складається зі згорток від conv1 до conv6b. У декодувальній (decoder) частині чергуються деконволюції (upconvN), звичайні згортки (iconvN, prN) та шари втрат. Ознаки з ранніх шарів конкатенують з ознаками вищих шарів.
- **DispNetCorr1D:** покращена версія з 1D-шаром кореляції, що виконує горизонтальне кореляційне зіставлення. Це обчислювально

ефективніше за 2D-кореляцію та дає змогу здійснювати тоншу дискретизацію на ширших діапазонах зміщень, що пояснюється тим, що одномірна природа задачі оцінювання диспаратності дозволяє обчислювати кореляції на тоншій сітці, ніж у FlowNet.

- **SceneFlowNet:** для спільної оцінки потоку сцени автори об'єднали попередньо навчені ваги FlowNet і DispNet в одну більшу мережу. Вона одночасно оцінює оптичний потік, поточну диспаратність і диспаратність наступного кадру; зміну диспаратності обчислюють із часових оцінок диспаратності.

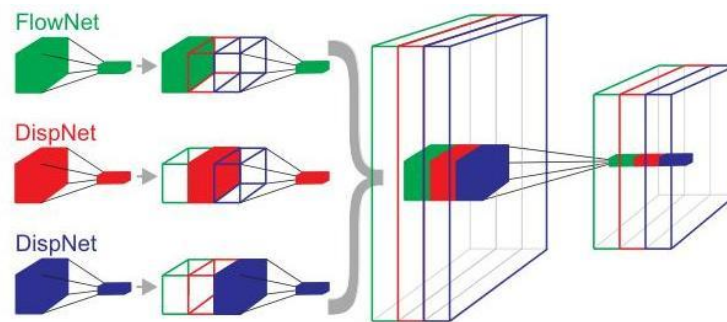


Рис. 1.25. Чергування ваг FlowNet (зелений) і двох DispNet (червоний та синій) у SceneFlowNet. Для кожного шару маски фільтрів створюють так: беруть ваги однієї мережі (ліворуч), а ваги інших мереж відповідно зануляють (посередині). Потім виходи кожної мережі конкатенують, утворюючи велику мережу з утрічі більшою кількістю входів і виходів (праворуч) [94]

У навчанні застосовано кілька ключових стратегій:

- наскрізна (end-to-end) оптимізація за допомогою оптимізатора Adam;
- прогресивне планування втрат: починаючи з втрат на грубій роздільності з поступовим урахуванням дрібніших деталей;
- розширене аугментування даних, зокрема просторові та колірні перетворення;
- багатомасштабна супервізія з функціями втрат на кількох рівнях роздільності.

1.5.2. LightStereo

Традиційні високоточні методи реконструкції карт глибини по стереоданим будують 4D-об'єм вартості (4D cost volume: висота \times ширина \times диспаратність \times ознаки) і використовують обчислювально витратні 3D згорткові нейронні мережі для агрегації вартості. Хоча ці підходи дають гарні якісні результати, вони споживають значні ресурси пам'яті й обчислень, а час інференсу часто перевищує 100 мс на кадр [96]. Попередні спроби полегшеної 2D-агрегації вартості, як-от MobilenetStereo-2D [97], супроводжувалися суттєвим зниженням точності.

Ключова проблема полягає в ефективній агрегації інформації про вартість уздовж виміру диспаратності, зберігаючи придатну обчислювальну ефективність. Більшість наявних полегшених методів або надто спрощують процес агрегації, або не здатні належно захопити багаті взаємозв'язки ознак, закодовані в каналному вимірі об'єму вартості.

LightStereo [96] пропонує 2D кодувально-декодувальну мережу, яка стратегічно зосереджується на підсиленні каналного виміру диспаратності в 3D-об'ємі вартості.

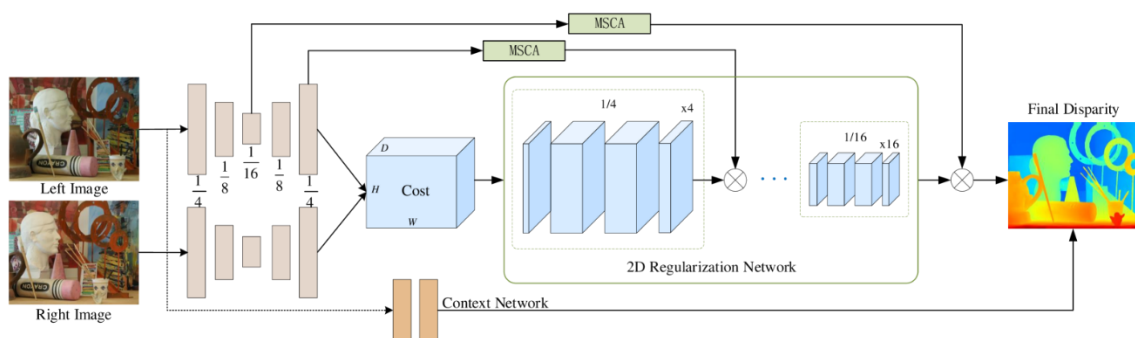


Рис. 1.26. Архітектура LightStereo підходу [96]

Архітектура складається з чотирьох основних компонентів:

Виділення ознак. Використовується бекбон MobileNetV2 [98] для вилучення багатомасштабних ознак на роздільностях 1/4, 1/8, 1/16 і 1/32. Прямі з'єднання (skip connections) та блоки апсемплінгу відновлюють ознаки до масштабу 1/4 для побудови об'єму вартості.

Побудова об'єму вартості. Замість 4D-об'єму LightStereo формує легший 3D кореляційний об'єм, обчислюючи подібність між лівими та правими картами ознак. Для кожного рівня диспаратності d кореляцію визначають як:

$$C_{cor}(d, h, w) = \frac{1}{C} \sum_{c=1}^C f_{l,4}(h, w) \cdot f_{r,4}(h, w - d), \quad (1.14)$$

де $f_{l,4}$ і $f_{r,4}$ – ліві та праві карти ознак у масштабі 1/4, а C – кількість каналів.

Агрегація вартості (cost aggregation – основна новизна [96]). Ключовий внесок полягає в стратегії агрегації вартості з використанням інверсних резидуальних блоків (Inverted Residual Blocks) з MobileNetV2 [98]. Ці блоки цілеспрямовано підсилюють каналний вимір диспаратності через:

- Розширення каналів: 1×1 згортка збільшує кількість каналів диспаратності;
- Глибинна обробка: 3×3 depthwise-згортки ефективно опрацьовують просторову інформацію;
- Проекцію каналів: ще одна 1×1 згортка зменшує розмірність із мінімальними втратами інформації;
- Skip-з'єднання: поліпшують протікання градієнтів, коли розмірності входу/виходу узгоджені.

Модуль багатомасштабної згорткової уваги (Multi-Scale Convolutional Attention, MSCA). Додатково метод включає модуль MSCA, який підсилює агрегацію, інтегруючи багатомасштабні ознаки з лівого зображення стереопари за допомогою ефективних «смугових» згорток (strip convolutions) з різними розмірами ядер (1×1 , 7×1 , 1×7 тощо).

Поєднання швидкодії (17 мс для LightStereo-S, при інференсі на RTX 3090 GPU) і точності робить можливим розгортання даного підходу на пристроях з обмеженими ресурсами для задач ДР.

1.5.3. AnyNet

У роботі [99] представлено AnyNet – архітектуру глибинного навчання, розроблену для подолання компромісу між швидкістю й точністю в оцінюванні

глибини зі стерео на кінцевих пристроях. На відміну від традиційних підходів, де потрібно обирати між швидкими, але неточними результатами, і повільними, але точними, AnyNet дозволяє динамічно запитувати результат у довільний момент інференсу, надаючи поступово уточнювані карти диспаратності зі зростанням доступного часу обчислень.

AnyNet використовує чотиристадійну архітектуру, що поступово уточнює передбачення диспаратності за допомогою багатомасштабної обробки та резидуального прогнозування. Ключова інновація [99] полягає в обчисленні початкової грубої карти диспаратності на дуже низькій роздільності, а далі – у прогнозуванні лише невеликих резидуальних поправок на вищих роздільностях, що різко зменшує обчислювальні витрати.

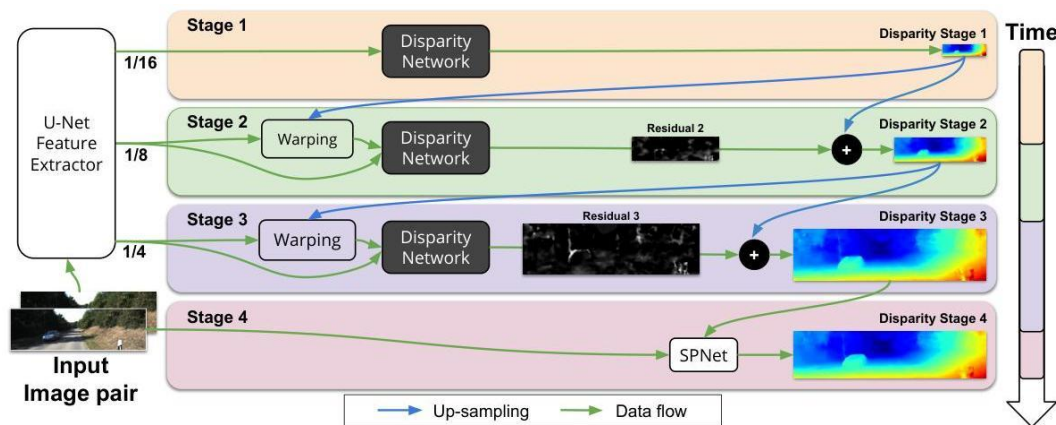


Рис. 1.27. Архітектура AnyNet із чотиристадійним процесом поступового уточнення [99]

Екстрактор ознак U-Net формує багатомасштабні ознаки; на стадіях 1–3 застосовуються мережі диспаратності з резидуальним передбаченням; на стадії 4 виконується доопрацювання просторовою розповсюджувальною мережею SPNet, що була запропонована в [100].

Архітектура підходу AnyNet містить чотири основні компоненти [99]:

U-Net екстрактор ознак. Спільна мережа виділення ознак опрацьовує ліві й праві вхідні зображення, формуючи багатомасштабні представлення ознак на роздільностях 1/16, 1/8 і 1/4 від оригіналу. Конструкція U-Net забезпечує ефективне обчислення ознак рівно тоді, коли вони потрібні наступним стадіям.

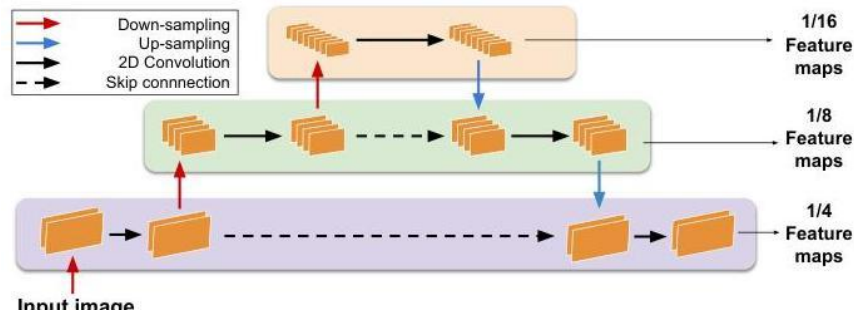


Рис. 1.28. Архітектура екстрактора ознак U-Net із поступовим зменшенням роздільності і skip-з'єднаннями для багатомасштабного виділення ознак [99]

Мережа оцінки диспаратності. Базовий модуль оцінювання диспаратності будує об'єм вартості (cost volume), обчислюючи подібність ознак між лівим і правим зображеннями для різних гіпотез диспаратності. Об'єм вартості уточнюється 3D-згортковими шарами, після чого застосовується *soft argmin* для регресії диспаратності.

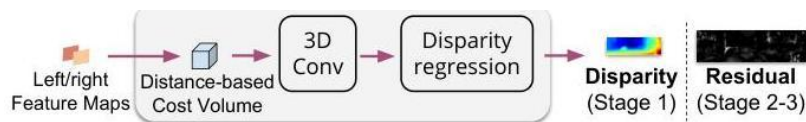


Рис. 1.29. Архітектура мережі диспаратності, що включає побудову об'єму вартості на основі відстані, 3D-згорткове уточнення та регресію диспаратності [99]

Багатоетапне резидуальне передбачення (основна новизна [99]):

- Етап 1: обчислює повну карту диспаратності на роздільності 1/16 від оригіналу із максимальним діапазоном диспаратності.
- Етапи 2-3: передбачають лише резидуальні поправки до збільшеної роздільності диспаратності з попередньої стадії; застосовується варпінг для узгодження ознак і потрібні лише малі діапазони пошуку диспаратності (± 2 пікселі).

SPNet (Spatial Propagation Network [100]). Підсумкова стадія виконує локальну фільтрацію для підвищення якості карти диспаратності та збереження дрібних деталей.

Процес навчання використовує спільну функцію втрат для всіх стадій із вагами $\lambda_1 = 1/4$, $\lambda_2 = 1/2$, $\lambda_3 = 1$, $\lambda_4 = 1$, заохочуючи кожну стадію продукувати оптимальні результати в межах свого обчислювального бюджету.

Час роботи моделі AnyNet на NVIDIA Jetson TX2 варіюється від 29 мс/кадр (1й етап, з середньою похибкою диспаратності 14%) до 100 мс/кадр (4й етап, з середньою похибкою диспаратності 6.2%).

1.5.4. Метод зниження вимог по пам'яті для доповнення карт глибини квантованою мережею

У роботі [101] представлено перше застосування змішаної точності квантизації до задачі доповнення глибини, що дає істотне зменшення використання пам'яті при збереженні високої якості реконструкції глибини для дуже розріджених ToF даних.

Дослідження [101] зосереджено на доповненні глибини з одного кадру для розрідженого об'єму ToF даних, де вхід складається з RGB-зображення та вкрай розрідженої карти глибини. На відміну від типових сценаріїв доповнення глибини з LiDAR, де частка валідних пікселів становить 4–5%, у цій [101] роботі розглядаються ToF-датчики, що дають лише 0.4–1.4% валідних вимірювань глибини.

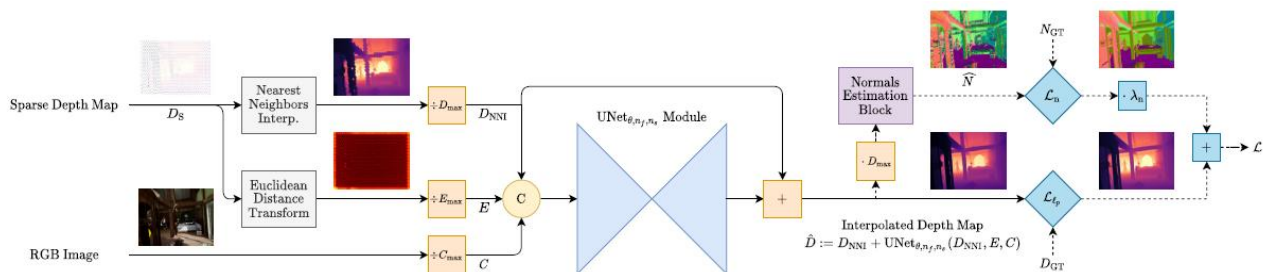


Рис. 1.30. Повне представлення методу: попередня обробка входів, архітектура U-Net і обчислення втрат [101]

Запропоноване рішення поєднує кілька ключових компонентів.

Попередня обробка входів. Застосовуються два кроки попередньої обробки запропоновані в методі D^3 [102]:

- Інтерполяція найближчим сусідом (Nearest Neighbors Interpolation, NNI): формує початкову щільну оцінку глибини, поширюючи найближче валідне значення глибини на невалідні пікселі;

- Евклідове перетворення відстані (Euclidean Distance Transform, EDT): генерує карту невизначеності, що відображає відстань до валідних вимірювань.

Отримані оброблені входи конкатенуються з RGB-зображенням, утворюючи багатоканальний вхід для мережі.

Базова архітектура моделі [101] представляє собою компактний U-Net-подібний енкодер-декодер із наступними характеристиками:

- 5 масштабів із 64 картами ознак на найвищому масштабі;
- стандартні 3×3 згортки з падінгом і кроком 1;
- шари апсемплінгу за допомогою NNI з подальшими згортками;
- резидуальний зв'язок, у межах якого мережа передбачає поправки до початкової оцінки NNI.

Ціль навчання поєднує точність за глибиною та геометричну узгодженість:

$$L(D) = L_{\ell_1}(D) + \lambda_n \cdot L_n(D), \quad (1.15)$$

де L_{ℓ_1} вимірює L1-відстань між прогнозованою та еталонною глибиною, а L_n оцінює косинусну подібність між прогнозованими та еталонними нормальми поверхні. Оптимальний ваговий коефіцієнт встановлено як $\lambda_n = 10^{-3}$ [101].

Ключовий технічний внесок [101] – застосування квантизації зі змішаною точністю в режимі навчання з урахуванням квантизації (QAT) до задачі доповнення глибини. Підхід дозволяє різним шарам мережі використовувати різні розрядності за умови дотримання загальних обмежень по використанню пам'яті.

Для поширення градієнтів використовується симетричний рівномірний квантизатор зі Straight-Through Estimator (STE). У [101] досліджено дві основні стратегії:

- Uniform Precision QAT: усі шари використовують однакову розрядність;
- Mixed Precision QAT: кожен шар навчається своїй оптимальній розрядності в межах обмежень по пам'яті.

Цільова функція для змішаної точності має вигляд:

$$L_{MP} = L(D) + \lambda_W \cdot P_W(S_W) + \lambda_A \cdot P_A(S_A), \quad (1.16)$$

де штрафні доданки забезпечують дотримання пам'яткових обмежень для ваг S_W і активацій S_A .

1.5.5. MobileViTv2

Візуальні трансформери (Vision Transformers, ViT) здійснили прорив у комп'ютерному зорі, досягнувши суттєвих на сьогодні результатів у низці задач, включаючи реконструкцію карт глибини та 3D реконструкцію середовища. Втім, їхнє практичне розгортання на кінцевих пристроях користувача суттєво обмежене обчислювальними витратами механізму багатоголової самоуваги (multi-headed self-attention, МНА). Традиційна МНА має квадратичну часову складність ($O(k^2)$) відносно кількості токенів k і значною мірою покладається на дорогі пакетні матричні множення, які погано узгоджуються з апаратними обмеженнями мобільних платформ.

У роботі [103] представлено MobileViTv2, що долає ці обмеження завдяки запропонованому механізму роздільної самоуваги (separable self-attention), який досягає лінійної складності ($O(k)$) та замінює обчислювально затратні операції на дружні до апаратури покомпонентні обчислення.

Ключова інновація полягає у роздільній самоувазі, яка концептуально переосмислює взаємодію токенів у процесі уваги. Замість попарних взаємодій між усіма токенами (як у традиційній МНА) запропоновано латентний токен, що виступає посередником для агрегування глобального контексту.

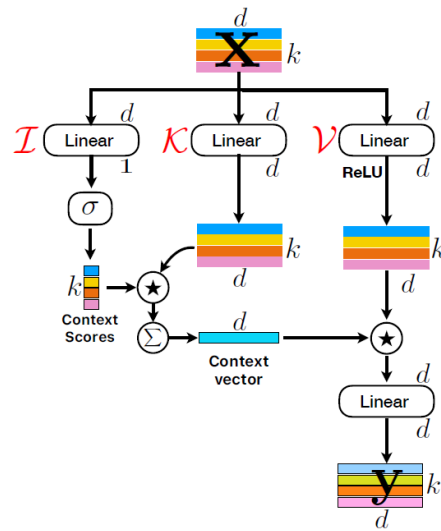


Рис. 1.31. Архітектура запропонованого методу з роздільною самоувагою та використанням покомпонентних операцій [103]

Механізм роздільної самоуваги працює у кілька кроків:

Обчислення контекстних оцінок. Для вхідних ознак x з k токенами розмірності d кожен токен проектується лінійним шаром ($W_I \in \mathbb{R}^d$) у скаляр. Ці скаляри відображають подібність між токеном і навчуваним латентним токеном. Далі застосовується *softmax* для перетворення до нормованих контекстних оцінок (C_S).

Глобальне агрегування контексту. Вхід проектується гілкою ключів ($W_K \in \mathbb{R}^{d \times d}$), утворюючи x_K . Глобальний контекстний вектор (C_V) обчислюється як зважена сума токенів (x_K) з вагами (C_S):

$$c_v = \sum_{i=1}^k c_s^i \cdot x_K^i \quad (1.17)$$

Розповсюдження контексту. Вхід також проектується гілкою значень ($W_V \in \mathbb{R}^{d \times d}$) з подальшою ReLU активацією, утворюючи x_V . Далі глобальний контекст C_V транслюється (broadcast) і застосовується до кожного токена через покомпонентне множення, ефективно повертаючи здобуту глобальну інформацію у представлення окремих токенів.

Ключова перевага полягає в тому, що весь процес спирається переважно на покомпонентні операції, а не на дорогі пакетні матричні множення, що робить його високоефективним на мобільному апаратному забезпеченні.

MobileViTv2 будується шляхом заміни модулів MHA у MobileViTv1 [104] запропонованою роздільною самоувагою. Архітектура має гібридний дизайн CNN-ViT: згорткові шари відповідають за вилучення локальних ознак, тоді як трансформерні блоки з роздільною самоувагою моделюють глобальні залежності.

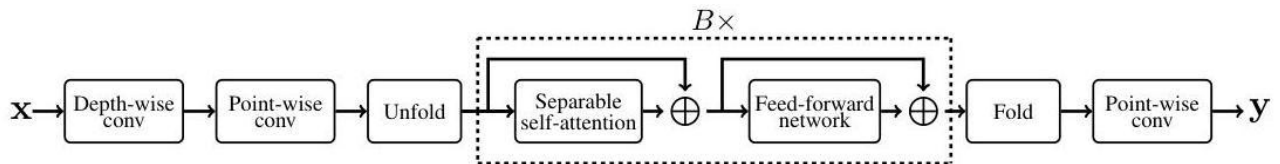


Рис. 1.32. Архітектура MobileViTv2, що ілюструє інтеграцію роздільної самоуваги в межах гібридної CNN-ViT архітектури [103]

Методика навчання включає розширені стратегії оптимізації:

- ImageNet-1k [105]: 300 епох, оптимізатор AdamW, косинусне планування швидкості навчання, розширене аугментування (RandAugment, CutMix, MixUp, Random Erase);
- Попереднє навчання: моделі попередньо навчаються на ImageNet-21k-P протягом 80 епох, далі – донавчання на ImageNet-1k;
- Багатомасштабна оцінка: донавчання на вищій роздільності 384×384 для підвищення якості.

1.6. Порівняльний аналіз методів та постановка задачі дослідження

З метою подальшого порівняльного аналізу розглянутих методів реконструкції карт глибини та 3D реконструкції середовища, проведемо їх групування:

1. Класичні швидкі методи з низькою затримкою:

- Логіка групування: класичні геометричні/фотометричні підходи з мінімальними вимогами до обчислень, орієнтовані на швидкість виконання.
- Методи:
 - Бінокулярна диспаратність (класичні алгоритми)
 - Паралакс руху (motion parallax)

- Відтворення форми за силуетом (Shape-from-Silhouette)
- Монокулярні евристики: лінійна перспектива, атмосферне розсіювання, розуміння глибини по затіненню, розмиття зображень

2. Класичні багатовидові підходи (офлайн реконструкція):

- Логіка групування: традиційний підхід реконструкції з високою точністю, але вищою обчислювальною вартістю – «положення камери (розріджений об'єм точок) → щільна карта глибини → 3D меш».
- Методи:
 - Структура з руху (Structure-from-Motion, SfM)
 - Багатовидове стерео співставлення (Multi-View Stereo, MVS)
 - Реконструкція поверхні
 - Наскрізний метод (end-to-end), який поєднує SfM, MVS та реконструкцію поверхні

3. Нейромережеві стерео підходи, орієнтовані на продуктивність:

- Логіка групування: стерео підходи для отримання диспаратності/глибини, сфокусовані на продуктивності та перенесенні на кінцеві пристрої.
- Методи:
 - DispNet
 - LightStereo
 - AnyNet

4. Нейромережеві MVS підходи з наскрізною (end-to-end) реконструкцією:

- Логіка групування: багатовидові DL-методи, які агрегують ознаки/глибину та часто навчаються наскрізним шляхом; менше ручних кроків, вища якість, більші обчислювальні ресурси.
- Методи:
 - HighRes-MVSNet
 - 3D-FHNet

- ATLAS
- SimpleRecon

5. *Методи реконструкції та рендерингу сцен:*

- Логіка групування: моделі представлення сцени для синтезу нових ракурсів.
- Методи:
 - Neural Radiance Fields (NeRF)
 - 3D Gaussian Splatting (3DGS)
 - NeRF похідні методи: BARF, MVSSNeRF, NerfingMVS, DDPNeRF, NSFF

6. *Дифузійні генеративні великі моделі для реконструкції щільної карти глибини та 3D реконструкції:*

- Логіка групування: дифузійні генеративні моделі для реконструкції карти глибини по монокулярним даним та 3D реконструкції; новітній клас із високими обчислювальними вимогами.
- Методи:
 - Marigold

7. *Пристроєорієнтовані підходи:*

- Логіка групування: ефективні нейромережеві методи (квантизація, мобільні трансформери) для виконанні в реальному часі на DSP/NPU.
- Методи:
 - Метод зниження вимог по пам'яті для доповнення карт глибини квантованою мережею (ідея методу може бути застосована для зниження вимог по пам'яті інших підходів)
 - MobileViTv2

Порівняльний аналіз складності реалізації та обчислювальної складності методів по запропонованим вище групам представлений у Табл. 1.3.

Табл. 1.3. Порівняння складності реалізації та обчислювальної складності методів

Група	Складність реалізації	Типовий час виконання / клас пристрою
1. Класичні швидкі методи	Низька – середня	В реальному часі на CPU / мобільному SoC
2. Класичні багатовидові підходи	Висока	Секунди–хвилини на сцену / GPU + CPU
3. Нейромережеві стерео підходи (ефективні)	Середня	В реальному часі на DSP/NPU/мобільних GPU
4. Нейромережеві MVS підходи	Висока – дуже висока	Секунди на кадр або хвилини на сцену / GPU
5. Методи реконструкції та рендерингу	Висока – дуже висока	Тренування: хвилини – години; рендер в реальному часі / GPU
6. Дифузійні генеративні великі моделі	Середня – висока	Сотні мс – секунди на зображення / GPU
7. Пристроєорієнтовані підходи	Середня	В реальному часі на DSP/NPU/мобільних GPU

Для коректної взаємодії в доповненій реальності недостатньо відтворити геометрію сцени, важливо також досягти субпіксельної точності реконструкції поверхонь, щоб віртуальні об'єкти перекривали реальні без візуальних артефактів, або навпаки, а відбивні та напівпрозорі поверхні наявні на сцені не порушували сприйняття. З метою забезпечення реалістичної взаємодії між користувачем, віртуальними та реальними об'єктами в сценаріях доповненої реальності, методи реконструкції карт глибини та 3D реконструкції мають забезпечувати: коректне розташування реконструйованих об'єктів, їх масштаби, оклюзії та колізії між доповненими та реальними об'єктами, коректне врахування напівпрозорих та відбивних поверхонь. Для цього результуюча 3D-модель повинна бути:

- Точною – глибина та нормалі повинні збігатися з реальними поверхнями. Невеликі похибки призводять до «плавання» віртуальних об'єктів у просторі та погіршує реалістичність сприйняття оклюзій і колізій між віртуальними та реальними об'єктами.

- Стійкою до змін та освітлення – реконструкція має враховувати наявність на сцені динамічних об'єктів, областей з сильними світловими відблисками, а також відбивні чи напівпрозорі поверхні.
- Енергоефективною – моделі повинні ефективно працювати на кінцевих пристроях користувача без зовнішнього енергоживлення, значного перегріву чи затримок.

Проведемо порівняльний аналіз розглянутих методів реконструкції карт глибини та 3D реконструкції середовища з точки зору їх спроможності генерувати нові види/ракурси середовища, їх стійкості до динамічного освітлення сцени, їх стійкості при наявності відбивних/напівпрозорих поверхонь, та їх здатності до адаптації під апаратні прискорювачі з обмеженою арифметикою (DSP/NPU). Результати порівняння представлені в Табл. 1.4.

Табл. 1.4. Здатність методів генерувати нові види, стійкість до складних сцен та спроможність до адаптації під DSP/NPU

Група	Генерація нових ракурсів	Стійкість до динамічного освітлення	Стійкість до відбивних / напівпрозорих поверхонь	Адаптація під DSP/NPU
1. Класичні швидкі методи	Здатність відсутня	Чутливі	Чутливі	Легко адаптуються
2. Класичні багатовидові підходи	Здатність відсутня	Чутливі	Чутливі	Складно адаптуються
3. Нейромережеві стерео підходи (ефективні)	Здатність відсутня	Помірно стійкі за узгоджених експозицій	Чутливі	Здатні до адаптації
4. Нейромережеві MVS підходи	Здатність відсутня	Помірно стійкі за узгоджених експозицій	Чутливі	Складно, або не адаптуються
5. Методи реконструкції та рендерингу	Спроможні до генерації	Чутливі	Помірно чутливі	Не адаптуються
6. Дифузійні генеративні великі моделі	Здатність відсутня	Чутливі	Помірно чутливі	Не адаптуються
7. Пристроєорієнтовані підходи	Здатність відсутня	Чутливі	Чутливі	Здатні до адаптації

Враховуючи обчислювальну складність розглянутих методів (Табл. 1.3), їх стійкість до складних умов середовища та спроможність до адаптації під апаратні прискорювачі з обмеженою арифметикою кінцевих пристроїв користувача (Табл.

1.4), можна зробити висновок, що підходи реконструкції карт глибини та 3D реконструкції потребують подальшого покращення та вдосконалення з метою забезпечення реалістичної взаємодії між користувачем, віртуальними та реальними об'єктами в сценаріях доповненої реальності.

Таким чином, **мета даного дослідження** полягає в підвищенні якості, стійкості та енергоефективності методів 3D реконструкції середовища, що виконуються на кінцевому пристрої користувача, для забезпечення реалістичної та надійної взаємодії користувача з віртуальними об'єктами в системах доповненої реальності.

Для досягнення мети дослідження сформульовані наступні **завдання дослідження**:

1. Виконати аналіз класичних та сучасних методів реконструкції карт глибини та 3D реконструкції середовища, визначити їхні обмеження для задач доповненої реальності на кінцевих пристроях користувача.
2. З метою врахування динамічних сцен та розширення існуючих наборів даних для задачі 3D реконструкції середовища, адаптувати технологію Neural Radiance Fields (NeRF) до умов динамічного освітлення.
3. Запропонувати метод обробки напівпрозорих та відбивних поверхонь для покращення реконструкції складних об'єктів.
4. Розробити метод ефективного прогнозування карт глибини високої точності з широким діапазоном глибини на кінцевих пристроях користувача та апаратних прискорювачах з обмеженою розрядністю.

1.7. Висновки до розділу

Стрімке розширення ринку споживання сценаріїв віртуальної та доповненої реальності на мобільних та носимих пристроях стимулюють розвиток методів реконструкції карт глибини та 3D реконструкції середовища, які повинні задовольняти суперечливі вимоги: забезпечувати високу геометричну точність та фотореалістичність реконструкції, адаптуватися до змін

освітлення й властивостей поверхонь, працювати в реальному часі на кінцевих пристроях користувача й витримувати обмеження обчислювальних ресурсів.

Розглянутий у розділі 1 стан галузі показав, що класичні та нейромережеві методи структури-з-руху та багатовидового стерео здатні реконструювати сцени зі статичним освітленням та дифузними поверхнями, але зазнають труднощів із динамічними джерелами світла та напівпрозорими/відбивними матеріалами. Крім того, більшість сучасних підходів 3D реконструкції середовища розраховані на обробку на сервері чи потужних GPU, тоді як носимі пристрої споживання сценаріїв доповненої реальності мають обмежені обчислювальні ресурси, пам'ять, енергоспоживання і часто використовують квантовані обчислення.

Отже, для покращення природності та імерсивності досвіду користувача в доповненій реальності, залишається низка відкритих питань та проблем, які потребують подальшого дослідження та удосконалення.

По-перше, необхідно розширювати та адаптувати існуючі набори даних до конкретних конфігурацій камер кінцевих пристроїв. Кожний носимий пристрій доповненої реальності має свої специфічні сенсори та їх параметри. Наявні набори даних, зазвичай не покривають необхідний кінцевий домен. Для покращення точності та надійності методів 3D реконструкції на кінцевому пристрої необхідно включати набори даних, що відповідають конкретним типами камер та їх характеристиками. Також, в наявних наборах даних погано присутні зразки з динамічним освітленням, наявністю відбивних та напівпрозорих поверхонь. Збір та анотація даних доволі тривала по часу та витратна процедура, отже і виникає необхідність адоптації існуючих наборів даних під кінцевий домен та врахування складних умов і поверхонь сцени.

По-друге, для повсякденного використання пристроїв доповненої реальності, важливо зосередити зусилля на покращенні точності реконструкції сцени в складних умовах. До таких сцен можна віднести динамічні сцени з великою кількістю рухомих об'єктів та сцени в яких присутні відбивні або напівпрозорі поверхні. Подолання цих викликів сприятиме покращенню якості

3D реконструкції середовища, отже матиме суттєвий вплив на якість сприйняття користувачем сценаріїв доповненої реальності.

Нарешті, пристрої споживання сценаріїв доповненої реальності, такі як смартфони та носимі окуляри ДР, обмежені з точки зору наявних обчислювальних ресурсів та ємності батареї. Якісні методи 3D реконструкції, зазвичай, мають високу обчислювальну складність і не відповідають вимогам обробки в реальному часі та вимогам по енергоспоживанню. Адаптація, портування, оптимізація методів 3D реконструкції та квантування відповідних моделей під кінцевий пристрій без суттєвого погіршення їх якості є критичною задачею для забезпечення ефективної роботи сценаріїв ДР, збільшення часу експлуатації носимих пристроїв та їх подальшої мініатюризації.

В даному розділі проведена постановка мети та завдань дослідження, орієнтована на підвищення якості, стійкості та енергоефективності методів 3D реконструкції середовища, що виконуються на кінцевому пристрої користувача, для забезпечення реалістичної та надійної взаємодії користувача з віртуальними об'єктами в системах доповненої реальності.

РОЗДІЛ 2: АДАПТАЦІЯ ТЕХНОЛОГІЇ NEURAL RADIANCE FIELDS (NERF) ДЛЯ ЗАДАЧІ 3D РЕКОНСТРУКЦІЇ СЦЕНИ В УМОВАХ ДИНАМІЧНОГО ОСВІТДЕННЯ

2.1. Проблематика динамічного освітлення в задачі 3D реконструкції

Тривимірна реконструкція сцени належить до фундаментальних задач комп'ютерного зору: йдеться про відновлення просторової структури середовища на підставі його двовимірних зображень. 3D-реконструкцію широко використовують у доповненій і віртуальній реальності. Зокрема, вона дає змогу коректно враховувати колізії та оклюзії між віртуальними об'єктами й фізичним світом, забезпечуючи природну та реалістичну взаємодію в ДР [22]. Складність задачі полягає в одночасному відтворенні глобальної геометрії сцени й тонких локальних деталей, що потребує значних обчислювальних ресурсів і великих обсягів даних. Високоякісні та достовірні дані є критично важливими для навчання глибоких моделей [106].

Навчальні набори зазвичай формують двома способами: через ручне збирання з подальшим анотуванням або за допомогою генерації синтетики. Ручний збір дорогий і трудомісткий. Такий підхід потребує синхронно отримувати зображення, еталонні значення глибини та точні зовнішні параметри камери, що вимагає спеціалізованих сенсорів – камер глибини та систем позиціонування. Повністю синтетичні дані не можуть повноцінно замінити реальні. Моделі, навчені лише на них, потребують додаткового донавчання на реальних прикладах з цільового домену, аби гарантувати належну якість роботи мережі в реальних умовах експлуатації [107].

Одним із перспективних підходів синтезу зображень останніх років є Neural Radiance Fields (NeRF) [81]. Це повнозв'язна мережа, яка, маючи обмежену кількість знімків сцени з відомими позами камер, навчається відтворювати нові ракурси, оптимізуючи базову неперервну об'ємну функцію на основі розрідженого набору вхідних даних [81]. NeRF придатний для генерування навчальних вибірок і аугментації даних для мереж передбачення

глибини. Наприклад, у [108] показано, що синтетичні зображення, згенеровані NeRF, підвищують точність регресії положення камери в задачах локалізації. Важливо й те, що така мережа здатна синтезувати не лише кольорові зображення, а й карти глибини, необхідні для оцінювання відстаней. Оскільки 3D-реконструкція має відтворювати реальну структуру середовища, надійні карти глибини є обов’язковими.

Під час навчання NeRF зазвичай мінімізує фотометричну функцію втрат – попиксельну відмінність між інтенсивностями еталонного та згенерованого зображень. Утім, за динамічних сцен або змінного освітлення така функція втрат виявляє обмеження. Залежність від інтенсивності порушує припущення про сталість яскравості, на якому ґрунтуються подібні фотометричні функції втрат. Аналогічні функції широко використовують у навчанні без учителя мереж передбачення глибини/відстані (наприклад, [109, 110]). Аналіз проблеми використання фотометричних функцій втрат на наборах даних з поганим чи динамічним освітленням проведено в [111].

В даному розділі запропоновано модифікацію методу NeRF з метою його адаптації для задачі реконструкції сцени та розширення наявних наборів даних в умовах динамічного освітлення. Апробацію запропонованого підходу проведено в роботі [112]. В роботі [22] детальніше розглянуто особливості моделювання середовища в умовах динамічного освітлення та наведено експериментальні результати, що підтверджують ефективність запропонованих змін.

2.2. Методологія

2.2.1. Обмеження фотометричної функції втрат NeRF підходу

Однією з характерних рис даних реальної сцени є часово змінна освітленість [22]. Натомість більшість наявних наборів даних охоплюють статичні сцени зі сталою освітленістю. Варіації освітлення можуть виникати через зовнішні джерела (сонце, фари автомобіля) або внаслідок налаштувань експозиції камери. Класичні моделі NeRF враховують положення та напрямки

вхідних променів, що певною мірою компенсує ефекти від відбивних поверхонь, однак часова змінна в них не моделюється. Для сцен із динамічним освітленням така залежність від часу є критично важливою. Без неї базові реалізації NeRF можуть формувати некоректні рендери сцени. Крім того, як зазначають автори [111], мережі передбачення глибини, які використовують фотометричні функції втрат, не здатні відновити 3D-структуру сцени на даних із поганою або змінною освітленістю. Відповідно, підходи на основі NeRF мають подібні обмеження, оскільки спираються на той самий тип функції втрат.

2.2.2. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Функція втрат за глибиною

В оригінальній моделі NeRF та її варіаціях одночасно передбачаються RGB-колір та об'ємна щільність σ , яку можна інтерпретувати як непрозорість. Використовуючи отриману щільність, можна відновлювати відстані до поверхонь об'єктів.

Стандартна процедура навчання NeRF не включає окремої оптимізації для передбачених карт глибини. Унаслідок цього оцінки відстаней погіршуються, що знижує якість 3D-реконструкції сцени. Якщо датасет зібрано за допомогою RGB-D камер, доступні (неповні) карти глибини можна використати для підвищення якості NeRF. Для цього вводимо додаткову функції втрат у вигляді середньоквадратичної похибки між еталонною та передбаченою картами відстаней:

$$L_D = \|D - \hat{D}\|_2^2, \quad (2.1)$$

де D – еталонні карти глибини, а \hat{D} – відповідні передбачення.

Тоді результуюча функція втрат має вигляд:

$$L = \sum_{i=1}^M \sum_u \|\hat{I}_i(u; w) - I_i(u)\|_2^2 + \|D - \hat{D}\|_2^2, \quad (2.2)$$

де M – кількість зображень, w – параметри моделі, що залежать від напрямів спостереження, u – координати пікселя, $\hat{I}_i(u; w)$ – синтезоване значення RGB-кольору в пікселі u для зображення i , $I_i(u)$ – еталонний RGB-колір у пікселі u

для зображення i , D – еталонна карта глибини, а \hat{D} – карта глибини, передбачена NeRF.

2.2.3. Модифікації оригінальної моделі NeRF для врахування умов динамічного освітлення. Час як додаткова вхідна змінна моделі NeRF

Щоб урахувати динамічне освітлення, пропонується додати час t як ще одну вхідну змінну мережі. Кожне значення t відповідає послідовному індексу навчального кадру. Оскільки вхідні дані є відеопослідовністю з фіксованою частотою кадрів (FPS), індекси часу та зображень зв'язані лінійно. Перед подачею до моделі змінну t нормалізуємо до інтервалу $[0,1]$ та піддаємо позиційному кодуванню. Параметр t додається лише на етапі навчання і не використовується на етапі виконання моделі. Пропонуються два способи інтеграції змінної t :

1. Додавання часу t до 3D-координат. Позначимо як (\vec{x}, t) .
2. Додавання t до напрямків вхідних променів. Позначимо як (\vec{d}, t) .

Відтак модифіковану модель NeRF можна подати так:

$$F_w : (\vec{x}, \vec{d}, t) \rightarrow (\vec{c}, \sigma) \quad (2.3)$$

де \vec{x} – просторове положення (x, y, z) , \vec{d} – напрям спостереження (θ, ϕ) , \vec{c} – RGB-колір для кожного пікселя, σ – щільність, що також використовується для розрахунку глибини по формулі (1.11).

2.3. Експерименти

2.3.1. Опис набору даних

Набір даних ScanNet [113] – це RGB-D датасет (кольорові зображення та відповідні карти глибини), що містить приблизно 2.5 мільйона зображень, зібраних у понад 1500 різних приміщеннях. Для кожної сцени надано внутрішні та зовнішні параметри камери (положення камери). У цій роботі використано дві

сцени ScanNet: scene0000_00 і scene0005_01. Через наявність відбивних поверхонь, а також часові зміни, освітленість у цих сценах залежить від кута спостереження (Рис. 2.1). Найімовірніше, така динаміка освітлення зумовлена автоматичною експозицією камери. Сцена scene0000_00 містить 5577 зображень, тоді як scene0005_01 – 1449 зображень. Для експериментів було відібрано лише підмножину scene0000_00 із 500 кадрів, спеціально обрану так, щоб включати зображення зі змінною освітленістю.

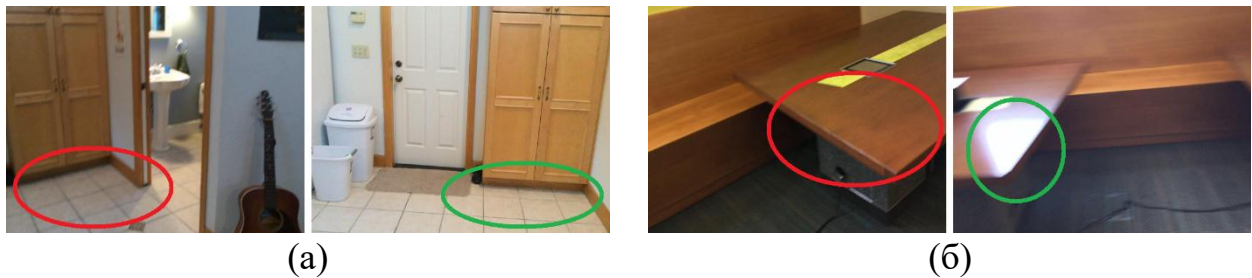


Рис. 2.1. Приклади сцен зі зміною освітлення поверхонь залежно від кута спостереження (а) та за наявності ефектів відбиття світла (б)

2.3.2. Опис експерименту. Метрики

Щоб продемонструвати вплив запропонованих змін, було проведено такі експерименти:

1. Навчання базової моделі NeRF.
2. Навчання NeRF із додатковою функцією втрат, що оптимізує карти глибин.
3. Навчання NeRF з введенням часової змінної як додаткового входного параметра.
4. Навчання NeRF із додатковими функціями втрат за глибиною та часовою змінною.

Архітектура в усіх випадках відповідає оригінальному NeRF [81]: MLP із 128 прихованими нейронами в кожному шарі. Для кожної сцени дані поділено у пропорції 90% / 10% на тренувальну й валідаційну вибірки. Під час навчання використано кожен другий кадр датасета. Зображення було масштабовано до 426×560 пікселів. Із кожного кадру випадково відбирали 2048 променів, уздовж кожного променя вибрано 128 точок для числового інтегрування. Діапазон

глибин для передбачення встановлено в межах 0 – 8 м. Для навчання був використаний оптимізатор Adam зі швидкістю навчання $1e-3$, що експоненційно знижувалася до $1e-4$; кожену модель тренували 200 тис. ітерацій.

Оцінювання здійснювали за двома метриками: середньою абсолютною похибкою (mean absolute error, MAE) та середньою абсолютною відносною похибкою (mean absolute relative error, MARE). Похибку рахували між передбаченою та еталонною глибиною. Обидві метрики є стандартними для порівняння карт глибин. Зокрема, у [89, 114] MARE використано як одну з величин для порівняння карт глибини.

MAE визначається наступним чином:

$$MAE = \frac{1}{N} \frac{1}{n} \frac{1}{m} \sum_{k=1}^N \sum_{i,j}^{n,m} |y_{k,i,j} - \hat{y}_{k,i,j}|, \quad (2.4)$$

де $y_{k,i,j}$ – еталонне значення глибини для (i, j) пікселя, $\hat{y}_{k,i,j}$ – передбачене значення глибини для (i, j) пікселя, N – кількість зображень у наборі даних, n – висота зображення, m – його ширина.

MARE обчислюється за формулою:

$$MARE = \frac{1}{N} \frac{1}{n} \frac{1}{m} \sum_{k=1}^N \sum_{i,j}^{n,m} \frac{|y_{k,i,j} - \hat{y}_{k,i,j}|}{y_{k,i,j}} * 100\%, \quad (2.5)$$

де $y_{k,i,j}$ – еталонне значення глибини для пікселя (i, j) , $\hat{y}_{k,i,j}$ – передбачене значення глибини для пікселя (i, j) , N – кількість зображень у наборі даних, n – висота зображення, m – його ширина.

Для підсумкового оцінювання було згенеровано всі кадри обох використаних сцен разом із відповідними картами глибин. MAE та MARE підраховано окремо для кожної моделі та кожного датасета. Під час підрахунку якості результатів, еталонні карти глибини відчищали від нульових значень, оскільки такі нулі спричинені специфікою застосованих сенсорів.

2.3.3. Результати експерименту

У Табл. 2.1 зведено результати розрахунку метрик для моделей, навчених у різних режимах. Для scene0000_00 навчання NeRF виконували без додаткової

функції втрат за глибиною. У базової моделі відносна похибка глибини становить 28.1%. Додавання часової змінної покращує якість моделі на 8–11%. Якісні результати для цього датасета наведено на Рис. 2.2. Верхній рядок відповідає RGB-зображенням, нижній – картам глибин. Усі карти глибини показано в єдиному інтервалі 1114 – 3737 мм, де темні відтінки відповідають меншим відстаням, світлі – більшим. Для кожної карти також вказано її власний діапазон у мм. Нульові значення еталонних глибинних карт (пропуски) вилучено з відображуваного інтервалу. Як видно з Рис. 2.2, введення часової змінної в NeRF підвищує якість як синтезованих RGB зображень (виділені області), так і карт глибин – прогнозований діапазон наближається до еталонного.

Табл. 2.1 Метрики для набору даних ScanNet

Набір даних	Режим навчання	MAE, м	MARE, %
scene0000_00	Базова модель	0.686	28.1
	(\vec{d}, t)	0.498	20
	(\vec{x}, t)	0.429	17.3
scene0005_01	Базова модель	0.577	30
	Функція втрат за глибиною	0.03	1.881
	Функція втрат за глибиною + (\vec{d}, t)	0.03	1.875
	Функція втрат за глибиною + (\vec{x}, t)	0.015	0.93

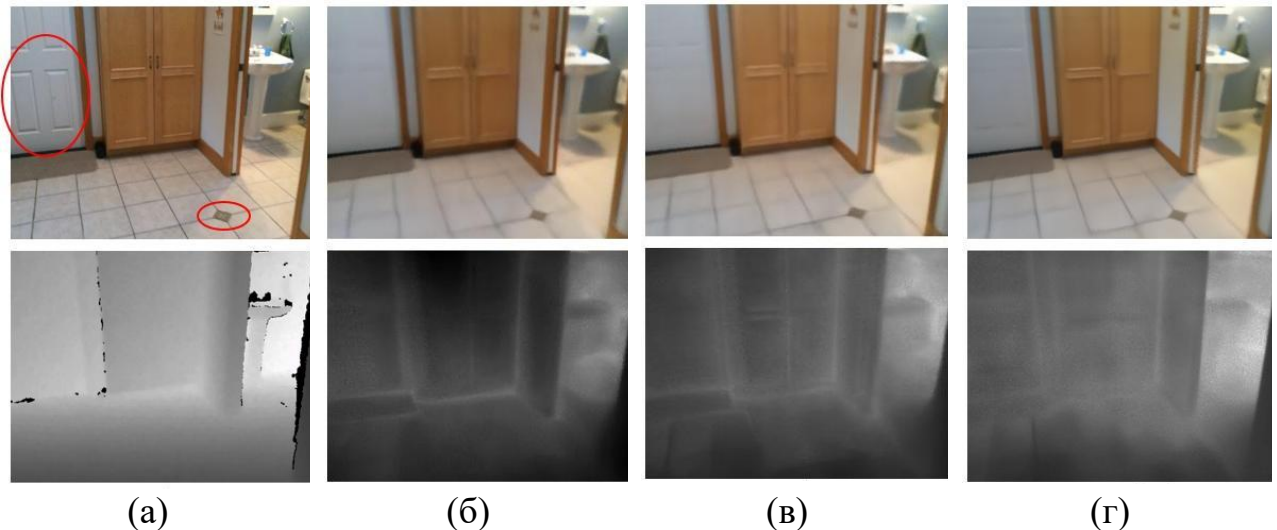


Рис. 2.2. Якісні результати порівняння для scene0000_00: еталонні дані, діапазон глибини 1605-3737 мм (а); NeRF, діапазон глибини 1114 - 2943 мм (б);
 NeRF + (\vec{d}, t) , діапазон глибини 1392 - 3660 мм (в);
 NeRF + (\vec{x}, t) , діапазон глибини 1510 - 3699 мм (г)

Для датасету scene0005_01 оцінено вплив обох модифікацій моделі: із функцією втрат за глибиною та з введенням додаткової часової змінної t . Використання функцією втрат за глибиною дозволяє зменшити відносну похибку з 30% до 1.88%. Додавання часової змінної додатково покращує якість як згенерованих кольорових зображень, так карт глибини. Найвиразніший ефект спостерігається, при додавання часової змінної до вхідних 3D-координат (\vec{x}, t) . У цьому режимі відносна похибка за глибиною становить 0.93%.

Якісні результати для scene0005_01 показано на Рис. 2.3: верхній рядок – RGB-зображення, нижній – карти глибин. Усі представлені карти глибини карти наведено в єдиному діапазоні 826 – 3472 мм: темні відтінки відповідають меншим глибинам, а світлі – більшим. Для кожної карти також подано її індивідуальний діапазон у мм. Нульові значення в еталонних глибинних картах вилучено з відображуваного інтервалу. Візуальне порівняння синтезованих зображень і глибин на Рис. 2.3 демонструє покращення якості генерації зображень завдяки запропонованим модифікаціям і дає змогу моделі точніше відтворювати окремі структури сцени (виділені області).

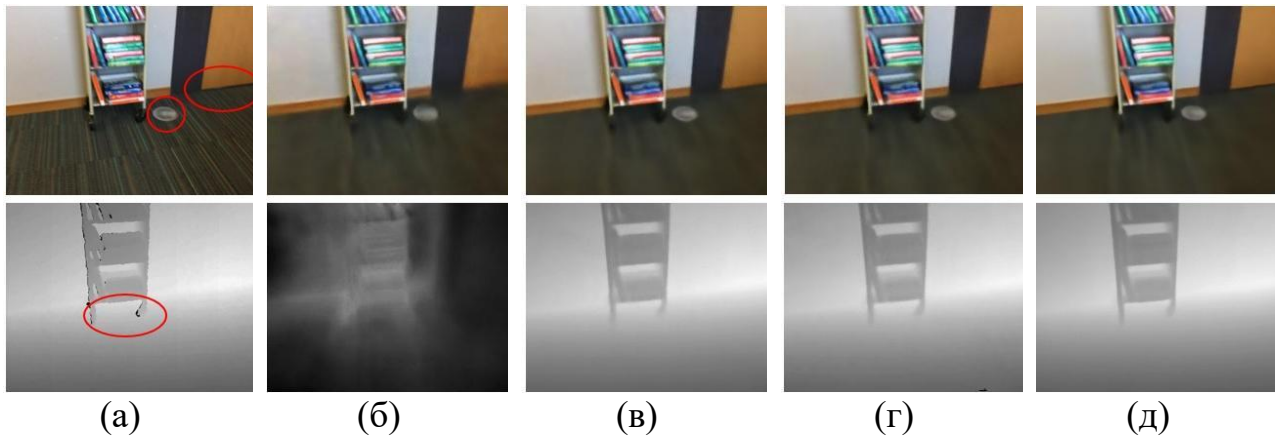


Рис. 2.3. Якісні результати порівняння для scene0005_01: еталонні дані, діапазон глибини 1501-2612 мм (а); NeRF, діапазон глибини 826-1909 мм (б); NeRF + функція втрат за глибиною, діапазон глибини 1443-2656 мм (в); NeRF + функція втрат за глибиною + (\vec{d}, t) , діапазон глибини 1451-2653 мм (г); NeRF + функція втрат за глибиною + (\vec{x}, t) , діапазон глибини 1502-2610 мм (д)

Додатково проведено якісний аналіз 3D-реконструкцій, отриманих різними конфігураціями моделі; підсумки представлено на Рис. 2.4. Інжекція часової змінної як додаткового вхідного аргументу до базової NeRF моделі дає змогу відтворити 3D меш з більшою кількістю структурних деталей і кращою відповідністю еталонній 3D-реконструкції. Зони помітних покращень позначено еліпсами.

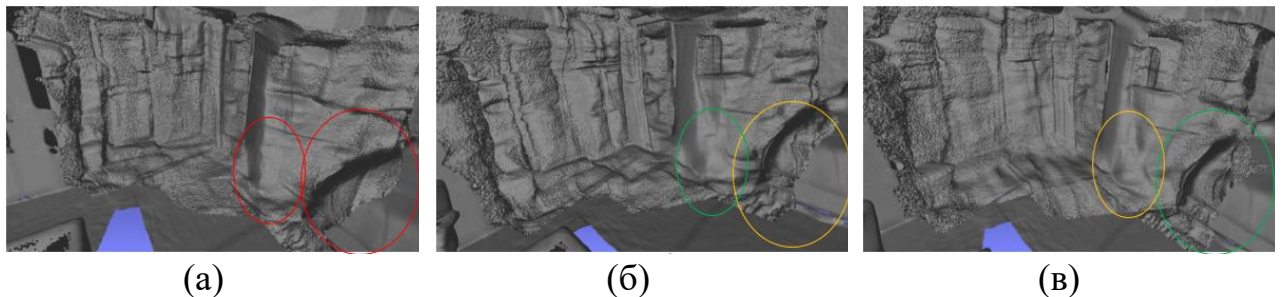


Рис. 2.4. Порівняння 3D реконструкції сцени для scene0000_00. Розширення вхідних параметрів часовою змінною покращує отриману структуру сцени.

Кольором виділені області з покращенням, де червоний колір позначає найгірший результат, зелений - найкращий.

а) NeRF; б) NeRF + (\vec{d}, t) ; в) NeRF + (\vec{x}, t) .

2.4. Висновки до розділу

У даному розділі досліджено вплив динамічного освітлення на якість представлення сцени NeRF моделлю. Такі зміни можуть спричинятися

джерелами світла зі змінною інтенсивністю (наприклад, сонячне світло в похмуру погоду), увімкненням/вимкненням ламп під час зйомки даних, або автоматичним налаштуванням експозиції камери. Продемонстровано, що стандартний підхід NeRF, який приймає лише просторову позицію та напрям кута спостереження, не здатен адекватно моделювати ці ефекти, що призводить до погіршення якості генерованих карт глибини. Щоб усунути цю ваду, запропоновано розширити вхід моделі часовою змінною та змінити функцію втрат. Ідею введення часу раніше застосовували для сцен із рухомими об'єктами; тут доведено доцільність того самого прийому для статичних сцен за умов динамічного освітлення. Експерименти на наборі даних ScanNet підтверджують, що включення часової змінної покращує фотореалістичність синтезованих зображень (зокрема для тонких структур) і зменшує відносну похибку глибини на 10–28%. Водночас метод успадковує типові обмеження NeRF, а саме, потребу навчати окрему модель під кожну сцену та значні часові витрати на навчання моделі. Додатково зберігається слабка придатність до реконструкції динамічних сцен із рухомими об'єктами. Результати даного дослідження можна використати для підвищення якості доповнення даних під час навчання моделей передбачення глибини, де критичною є точність і візуальна якість як зображень, так і карт глибини.

РОЗДІЛ 3: РЕКОНСТРУКЦІЯ КАРТ ГЛИБИНИ ВРАХОВУЮЧИ НАПІВПРОЗОРИ ТА ВІДБИВНІ ПОВЕРХІ

3.1. Проблематика впливу напівпрозорих та відбивних поверхонь на реконструкцію карт глибини

Тривимірна реконструкція середовища та оцінка глибини сцени є одними з ключових задач комп'ютерного зору, що знаходять застосування у різноманітних сферах – від доповненої і віртуальної реальності до робототехніки та автономного водіння. Для побудови адекватної 3D-моделі навколишнього світу системі зору необхідно точно визначати відстань до об'єктів на зображеннях, тобто будувати карти глибини. Карта глибини – це зображення, де кожному пікселю сцени поставлено у відповідність значення глибини (відстані до камери). Якісне відновлення глибини сцени є критично важливим для коректного просторового сприйняття: наприклад, для розміщення віртуальних об'єктів у ДР, для навігації роботів у просторі чи для виявлення перешкод в автомобілях з автономним керуванням.

Разом із цим, сприйняття глибини сцени стикається з цілою низкою проблем в реальних умовах. Зокрема, існуючі методи визначення глибини часто припускають, що поверхні в сцені мають дифузні, або ламбертівські властивості відбиття світла. Проте в навколишньому середовищі широко представлені дзеркальні (відбивні) поверхні – наприклад, дзеркала, поліровані металеві або пластикові об'єкти. Також в навколишньому середовищі широко розповсюджені прозорі чи напівпрозорі об'єкти – такі як скляні вікна, екрани тощо. Наявність таких об'єктів суттєво ускладнює оцінку глибини. Фотометричні припущення алгоритмів стереозору чи монокулярної оцінки глибини порушуються, оскільки зображення цих областей формуються відбитим або пройденим світлом [115]. В результаті стандартні підходи можуть давати некоректні або неповні результати глибини на відбивних та напівпрозорих ділянках сцени.

Для подальшого аналізу проблематики та постановки задачі зазначимо базову нотацію та опис камери:

- Маємо камеру з центром o і напрямком променя $d(u)$ для пікселя $u \in \Omega \in \mathbb{R}^2$.
- Промінь з камери: $r_u(t) = o + td(u)$, $t > 0$, де t – параметр уздовж променя, що демонструє наскільки далеко від камери ми рухаємось у напрямку $d(u)$.
- Реальна сцена: множина точок $S \subset \mathbb{R}^3$, що існують у фізичному просторі.
- Глибина реального об'єкту (евклідова або вздовж променя) для пікселя u : $D^*(u) = \inf \{t > 0 \mid r_u(t) \in S\}$.

Нехай промінь r_u спершу перетинає напівпрозору поверхню в точці $x_{surf} = r_u(t_{surf})$, далі об'єкт за поверхнею у точці $x_{behind} = r_u(t_{behind})$, де $0 < t_{surf} < t_{behind}$. Тоді інтенсивність:

$$I(u) = (1 - \tau(u))C_{surf}(u) + \tau(u)C_{behind}(u), \quad (3.1)$$

де $\tau(u) \in (0,1)$ – ефективна пропускна здатність у напрямку променя (функція матеріалу, кута падіння, товщини), C_{surf} – внесок від самої поверхні, C_{behind} – від об'єкта за поверхнею.

Особливо критичними є такі випадки [20]:

1. Напівпрозорі об'єкти. Якщо в сцені присутнє прозоре або напівпрозоре тіло (наприклад, скло), то традиційні алгоритми можуть взагалі не виявити його як окремий об'єкт за даними глибини. На Рис. 3.1 патенту ілюстровано ситуацію, коли скляна перегородка залишається нерозпізнаною та відсутня на карті глибини.

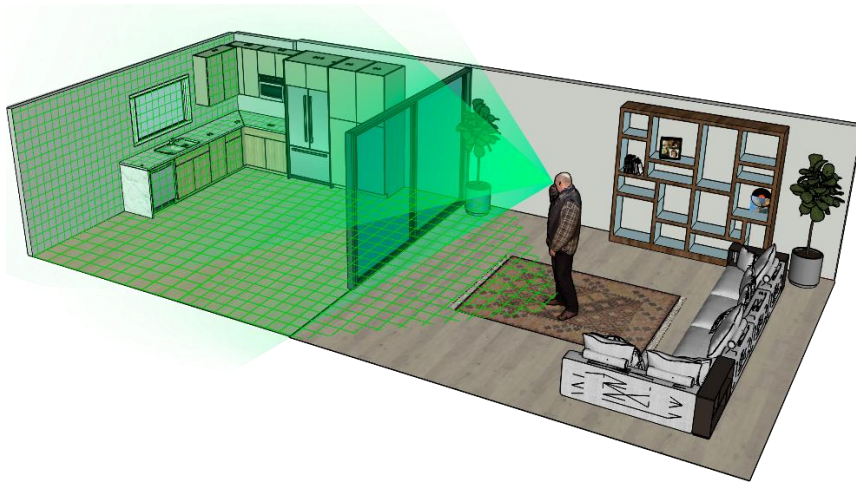


Рис. 3.1. Ілюстрація проблеми оцінки карти глибини при наявній на сцені напівпрозорій поверхні. Відсутня глибина до напівпрозорої поверхні [20]

Метод оцінює $\hat{D}(u) = t_{\text{behind}}$, та відображає на карті глибини точки $r_u(\hat{D}(u)) \in S_{\text{behind}}$, що належать об'єктам за напівпрозорою поверхнею. Але метод втрачає точки глибини самої напівпрозорої поверхні $r_u(\hat{D}(u)) \notin S_{\text{surf}}$.

З іншого боку, методи сприйняття глибини можуть не бачити крізь напівпрозору поверхню, та не зафіксувати об'єкти, що знаходяться безпосередньо за нею. На Рис. 3.2 показано, що предмети позаду напівпрозорої перешкоди не потрапили до реконструйованого 3D-середовища.

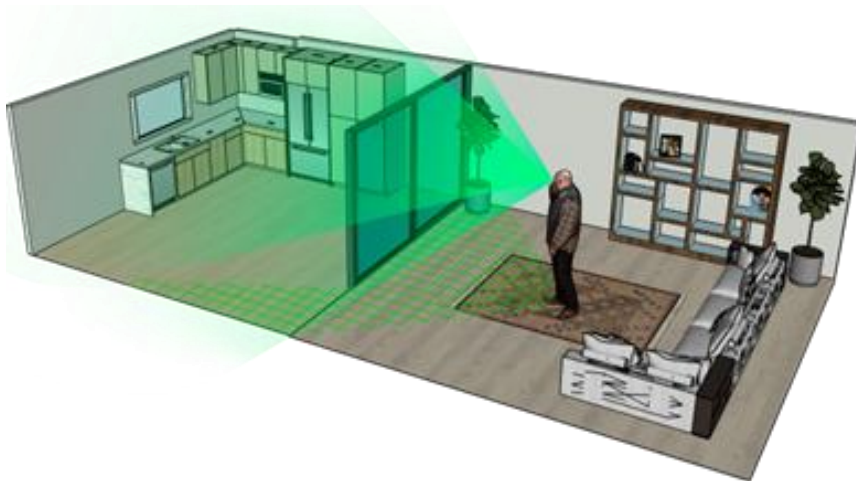


Рис. 3.2. Ілюстрація проблеми оцінки карти глибини при наявній на сцені напівпрозорій поверхні. Відсутня глибина до об'єктів за поверхнею [20]

Метод оцінює $\hat{D}(u) = t_{\text{surf}}$, та відображає на карті глибини точки $r_u(\hat{D}(u)) \in S_{\text{surf}}$, що належать напівпрозорій поверхні. Але метод втрачає точки глибини об'єктів за напівпрозорою поверхнею $r_u(\hat{D}(u)) \notin S_{\text{behind}}$.

Таким чином, при наявній напівпрозорій поверхні на сцені, реконструкція 3D сцени може втрачати важливі елементи: або саму напівпрозору поверхню поверхня, або те, що за нею.

2. Дзеркальні поверхні. У випадку відбивних об'єктів ситуація протилежна – алгоритми сприйняття глибини помилково знаходять в дзеркалі уявні об'єкти. На Рис. 3.3 демонструється, як дзеркало відображає сцену, і підхід сприйняття глибини інтерпретує відбиті об'єкти як реальні, що помилково розміщуються позаду дзеркальної поверхні. Наприклад, людина, що знаходиться перед дзеркалом, може бути відображена в ньому, і алгоритм оцінки карт глибини розташовує її позаду дзеркала, ніби за стіною. Такі спотворення призводять до некоректної 3D реконструкції. Відбивні поверхні маскують реальну геометрію сцени або викривляють її відтворення.

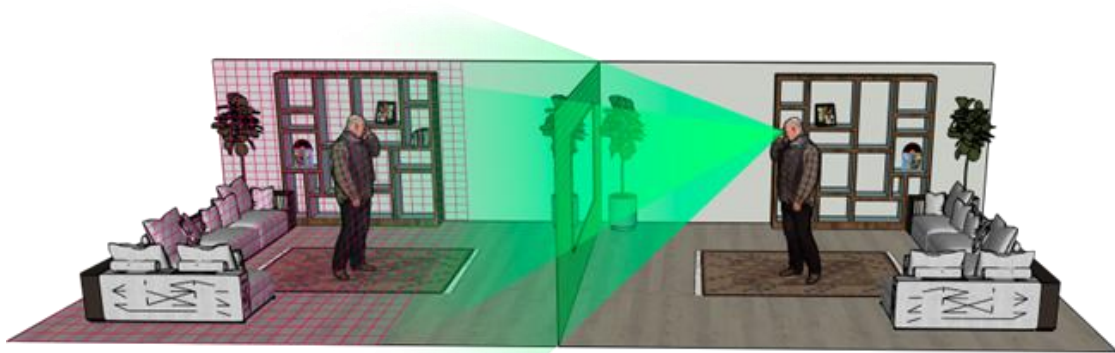


Рис. 3.3. Ілюстрація проблеми оцінки карти глибини при наявній на сцені відбивної поверхні. На карті глибини з'являються об'єкти за відбивною поверхнею [20]

Для дзеркальної площини $\Pi = \{x \mid n^T x = d\}$ відбитий напрямок:

$$d_r(u) = d(u) - 2(d(u)^T n)n \quad (3.2)$$

Точка реального об'єкта $p \in S$ має відбиту позицію:

$$p_{mir} = p - 2(n^T p - d)n \quad (3.3)$$

За наявності ідеального відбиття, інтенсивність уздовж r_u еквівалентна спостереженню p_{mir} у напрямку $d(u)$. Алгоритм, що не розпізнав відбивну поверхню, оцінює $\hat{D}(u) = t_{mir}$, та відображає на карті глибини точки $r_u(\hat{D}(u)) \in p_{mir}$, тобто включає у карту глибини уявні точки, які належать реальним об'єктам $p \in S$

, але розміщені по інший бік відбивної площини Π . Алгоритм втрачає точки самої відбивної поверхні Π : $r_u(\hat{D}(u)) \notin S$.

Покращення методів побудови карт глибини для сцен з відбивними та напівпрозорими поверхнями є актуальним завданням. У патенті [20] запропоновано новий метод генерації карт глибини, що дозволяє врахувати присутність таких складних поверхонь і підвищити повноту та точність 3D реконструкції сцени.

Більшість сучасних алгоритмів отримання глибини можна поділити на дві великі категорії: стереоскопічні методи та методи монокулярної оцінки глибини на основі навчання. Стерео-методи використовують пару (або більше) зображень однієї сцени, зроблених з різних ракурсів, і визначають глибину шляхом пошуку відповідних точок (пікселів) між зображеннями – стереовідповідності. Класичні підходи включають алгоритми кореляції, динамічного програмування, напівглобального узгодження тощо, які будують карти диспаратності (карти різниці координат) і перетворюють їх у глибину. Проте стереоалгоритми припускають, що сцена спостерігається з різних точок під приблизно однаковими умовами освітлення і відбиття. Відбивні та прозорі поверхні порушують це припущення: вони не мають стабільних візуальних ознак, за якими можна знайти відповідність на двох зображеннях, оскільки кожна камера бачить в таких областях зовсім інше зображення (іншу частину сцени або спотворене відбиття). Це призводить до того, що стандартні стерео-методи не можуть коректно визначити диспаратність на дзеркальних ділянках, або ж визначають їх невірно, наприклад, прив'язуючи їх до фону. Як наслідок, карта глибини від стереосистеми має пропуски або артефакти в зонах відбивних, або напівпрозорих об'єктів.

Монокулярна оцінка глибини за допомогою глибоких нейронних мереж останнім часом досягла значного прогресу. Згорткові нейронні мережі (CNN) навчилися прогнозувати глибину зі звичайного одиночного RGB-зображення, використовуючи накопичений досвід (дані) про типові сцени. Багато сучасних досліджень орієнтовані на покращення точності монокулярної глибини завдяки

архітектурам CNN, вдосконаленим функціям втрат та використанню величезних навчальних наборів даних. CNN добре виділяють локальні особливості зображення і можуть на основі текстурних та контекстних підказок оцінити глибину різних областей сцени. Однак, залишається фундаментальне обмеження – відсутність прямих даних про глибину робить задачу неоднозначною, і модель часто опирається на припущення про навколишній світ. Випадки, які вибиваються із звичних шаблонів, викликають у неї помилки. Так, дзеркальні чи напівпрозорі поверхні є типовою проблемою – вони порушують ламбертівське припущення про постійність візуальних властивостей поверхні при зміні ракурсу [115]. Якщо модель навчалась переважно на сценах без дзеркал, вона може просто не зрозуміти, що робити з їх відображеннями. Навіть сучасні навчальні підходи, такі як Self-Supervised Monocular Depth, потерпають від цієї проблеми, адже алгоритм, навчаючись на послідовностях кадрів, намагається мінімізувати фотометричну невідповідність між кадрами – а на відбивних поверхнях така невідповідність неминуча і збиває алгоритм. У результаті, як показано у дослідженні [115], на відображеннях глибини визначаються дуже неточно. З іншого боку, архітектури з трансформерами нещодавно продемонстрували певні переваги в задачі оцінки глибини завдяки здатності враховувати глобальний контекст зображення. Зокрема, моделі на основі Vision Transformer (ViT) досягли високої точності на деяких наборах даних, перевершивши традиційні CNN [116]. Трансформери краще розпізнають складні структури текстур, що може допомагати визначати глибину за непрямыми ознаками. Однак, як показано у порівняльних роботах, трансформерні моделі можуть втрачати безперервність глибинних градієнтів та вимагати значно більше даних для навчання [116, 117]. В контексті відбивних поверхонь, навіть найдосконаліші архітектури не гарантують правильного результату, якщо явно не передбачено механізму обробки відбиттів. Отже, особливість проблеми полягає в тому, що ані стереосистеми, ані сучасні моно-мережі не забезпечують коректної оцінки карт глибини при наявності дзеркальних та напівпрозорих об'єктів. Це обмежує

застосування 3D-реконструкції в реальних середовищах, де такі об'єкти зустрічаються повсюдно.

Для подолання згаданих обмежень, запропоновано відмовитися від традиційної концепції єдиного значення глибини на точку зображення. У випадку прозорих та дзеркальних поверхонь одна точка зображення фактично відповідає двом фізичним об'єктам: самій поверхні (склу або дзеркалу) та об'єкту за нею або відображенню, що також спостерігається в цій точці. Тому запропоновано формувати розширену карту глибини, яка містить декілька шарів глибини для таких ділянок [20]. Зокрема, замість неоднозначних значень карти глибини на відбивних об'єктах, метод дозволяє явно отримати два значення: до поверхні дзеркала/скла та до реального об'єкта, що за ним стоїть або у ньому відбивається [20]. Ця концепція реалізована шляхом поєднання двох часткових карт глибини, отриманих різними способами. Важливо, що такий підхід по суті додає сцені вимір напівпрозорості: реконструйоване 3D-середовище включає як самі складні об'єкти, так і те, що крізь них видно, що наближає модель до реальної композиції сцени [20].

Отже, постановка задачі, яку вирішує запропонований метод – навчитися генерувати достовірні карти глибини в присутності прозорих та дзеркальних об'єктів [20]. Це потребує вирішення наступних підзадач:

1. виявлення на зображенні областей, що відповідають відбивним або напівпрозорим поверхням;
2. окремого визначення глибини для цих поверхонь та для об'єктів, які вони відображають або приховують;
3. інтеграції цієї інформації в єдину структуру, придатну для подальшої 3D-реконструкції (наприклад, у вигляді сукупності полігональних сіток чи глибини на кількох шарах).

3.2. Метод реконструкції карт глибини, що враховує напівпрозорі та відбивні поверхні

Запропонований метод реалізує згадану вище ідею через послідовність етапів обробки зображень [20], яка продемонстрована на Рис. 3.4. Система, що виконує задачу оцінки карти глибини та реконструкції, розглядається як електронний пристрій, наприклад: смартфон, окуляри ДР чи інший комп'ютерний пристрій з вбудованою камерою та засобами обробки зображень. На вході метод отримує принаймні одне зображення цільової сцени, яка може містити відбивні або напівпрозорі об'єкти. Вхідні дані не обмежуються одним кадром – це може бути послідовність RGB-кадрів (відео фрагмент), стереопара або комбінація зображень, а також можуть бути додаткові дані з RGB-D камер. Метод гнучко пристосований до різних джерел: якщо є RGB-D камера, то первинну глибину можна взяти з неї; якщо ж ні – використовуються тільки RGB-кадри, а карти глибини реконструюються за допомогою нейронних мереж. Далі метод умовно поділяє процес на дві паралельні гілки: одна відповідає за глибину до кожного самостійного об'єкта, інша – за глибину до об'єктів, що проглядаються через напівпрозорі об'єкти, або відбиваються у дзеркальних об'єктах.

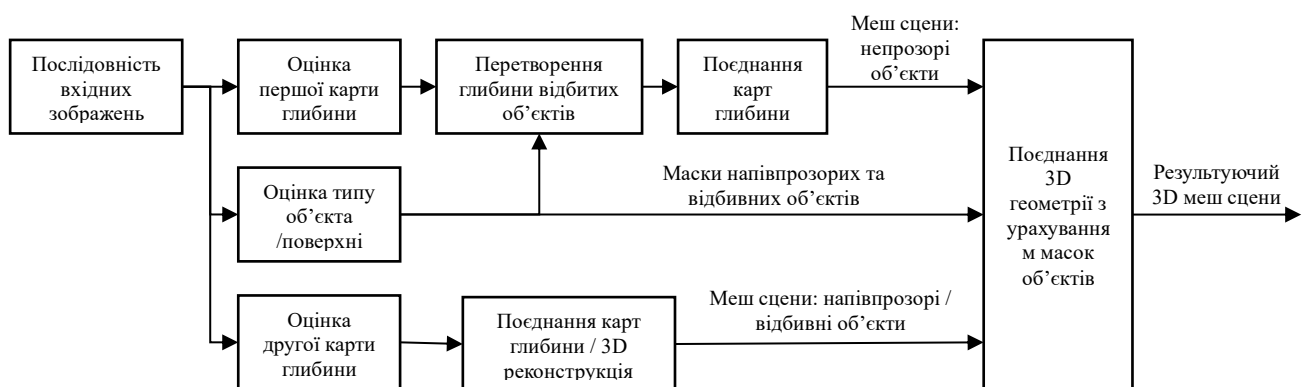


Рис. 3.4. Спрощена блок-схема методу реконструкції сцени за наявності відбивних та напівпрозорих об'єктів

Маючи зображення I (або відеопослідовність) потрібно оцінити:

1. Відстань до реальних точок поверхонь сцени S (включно з напівпрозорими/відбивними).

2. Відстань до об'єктів за напівпрозорими поверхнями або об'єктів, що відбилися.
3. Тип поверхні в кожному пікселі.

Побудувати двоканальну карту глибин:

$$D(u) = (D_{surf}(u), D_{obj}(u)), \quad (3.4)$$

де D_{surf} – глибина до самої (напівпрозорої/відбивної) поверхні, а D_{obj} – глибина до об'єкта позаду (для напівпрозорих) чи до відбитого об'єкта (для відбивних).

Як показано на Рис. 3.4, основою методу є побудова двох проміжних карт глибини ($\hat{D}_{surf}(u)$ і $\hat{D}_{obj}(u)$) та їхнє наступне злиття. Розглянемо детальніше основні етапи методу:

1) Визначення глибини прихованих об'єктів. На першому етапі із вхідного зображення (або послідовності зображень) алгоритм отримує первинну оцінку глибини сцени, ігноруючи вплив відбивних поверхонь. Система намагається зазирнути крізь відбивні та напівпрозорі поверхні. Для цього формується перша карта глибини, що містить значення глибини для непрозорих об'єктів, навіть якщо вони спостерігаються шляхом відбиття ($r_u(\hat{D}(u)) \in p_{mir}$) або крізь напівпрозорий матеріал $r_u(\hat{D}(u)) \in S_{behind}$. Іншими словами, на цій карті глибини відбивні та напівпрозорі поверхні як такі не представлені – натомість в тих ділянках позначена глибина об'єктів, що відбиваються або знаходяться за склом. Така карта може бути отримана різними шляхами. Якщо, наприклад, система має стереопару або датчик глибини, можна безпосередньо отримати глибину сцени. Така карта глибини може не міститиме коректних даних на відбивних поверхнях – ці області, ймовірно, з'являться як шуми чи порожнечі). У випадку відсутності прямого сенсора глибини, пропонується використати алгоритми оцінки глибини із зображень. При цьому, звичайний підхід не враховує віддзеркалення, тому першу карту глибину можна отримати із застосуванням наявних глибинних методів прогнозування з припущенням ламбертівських поверхонь. Зокрема, може бути використана нейронна мережа для регресії глибини – тренована модель, яка на основі одного або декількох кадрів видає карту глибини. При

цьому модель, як правило, інтерпретує відблиски як частину дальнього фону або ігнорує. Так чи інакше, результатом першої гілки є карта глибини, де для кожного видимого на зображенні непрозорого об'єкта задано його глибину. Якщо об'єкт видно напряду – це просто його відстань; якщо об'єкт видно тільки у дзеркалі – на цій карті все одно повинна з'явитися його відстань у відповідному місці, де було відображення. Аналогічно, об'єкти за напівпрозорими поверхнями мають бути присутніми на цій глибинній карті, ніби напівпрозора поверхня відсутня. Для досягнення цього метод включає спеціальний підпроцес: виявлення відбитих/прихованих об'єктів. Семантична сегментація сцени дозволяє зрозуміти які наявні об'єкти є відбивними або напівпрозорими, а також які об'єкти є їхніми відбиттями: $s(u) \in \{opaque, translucent, reflective\}$. Застосовуючи методи навчання, наприклад, згадану регресійну мережу або інший алгоритм, пристрій ідентифікує на зображенні дзеркальні області та знаходить об'єкти, що в них відображені [20]. Для відображених об'єктів виконується оцінка глибини за непрямыми ознаками, враховуючи положення та кривизну відбивної поверхні, тощо. Після цього значення глибини вписуються в першу карту глибини на місця відбивних областей. Таким чином, перша карта доповнюється глибинами прихованих реальних об'єктів. Умовно кажучи, після першого кроку ми маємо карту глибини сцени, очищену від ефектів відбиття – в ній присутні глибини тільки фізичних непрозорих об'єктів з правильними значеннями, навіть якщо ті були видні лише через дзеркало або скло.

2) Визначення глибини відбивних та прозорих поверхонь. Другою паралельною гілкою методу є побудова другої карти глибини, яка містить вже значення глибин самих складних поверхонь – дзеркал, блискучих площин, скла тощо [20]. Цей крок виявляє де в просторі знаходиться відбивна або напівпрозора поверхня ($r_u(\hat{D}(u)) \in S_{surf}$). Для кожного відбивного, чи напівпрозорого фрагменту сцени визначається його власна глибина, тобто відстань від камери до цієї поверхні. Звичайні методи оцінки кари глибини якраз і не повертають коректну відстань до таких регіонів, тому необхідно залучати додаткові засоби. Патент пропонує два підходи, які не є взаємовиключними: або використати стереозір,

або застосувати алгоритм машинного навчання. У першому випадку, якщо доступні два ракурси сцени, можна спробувати знайти відповідності на відбивній поверхні (це можливо, якщо, наприклад, на ній є часткові забруднення чи особливості, або якщо використовується активний сенсор визначення карти глибини. У другому випадку – тренувати нейронну мережу спеціально для оцінки глибини дзеркальних поверхонь. Важливим також є виявлення, які області зображення відповідають відбивним та напівпрозорим поверхням. Це завдання вже частково вирішене на попередньому етапі через семантичне розуміння сцени – відомі маски тих областей, що є відбивними чи напівпрозорими. Отримавши ці маски, алгоритм призначає кожній такій області певне значення глибини. Існують різні ознаки, за якими можна оцінити глибину дзеркальної поверхні: наприклад, геометричні співвідношення між положенням реального об'єкта та його відбиттям (якщо відомі хоча б приблизні відстані до об'єкта і відображення, можна геометрично вивести позицію дзеркала. Також можна використати припущення про безперервність глибини. Глибина поверхні дзеркала може бути оцінена на основі глибини сусідніх непрозорих поверхонь, які до нього прилягають. У будь-якому разі, другий результат – це карта глибини, де на місці кожної відбивної чи напівпрозорої поверхні зазначена глибина саме до точки цієї площини. На відміну від першої карти глибини, де присутні відстані до відбитих об'єктів та об'єктів за напівпрозорими поверхнями, друга карта глибини вказує відстані до відбивних та напівпрозорих поверхонь, ігноруючи відбиті чи приховані об'єкти. Таким чином, дві карти глибини доповнюють одна одну.

3) Злиття карт та формування результуючого представлення сцени. На завершальному етапі запропонований метод об'єднує інформацію з двох джерел, отримуючи фінальну карту глибини сцени [20]. Це злиття не тривіальне, оскільки тепер для деяких пікселів, що відповідають дзеркальним чи прозорим областям, ми маємо два різні значення глибини від першої карти (глибина відбитого чи прихованого об'єкта), інше від другої (глибина поверхні). Запатентований метод пропонує кілька способів інтеграції. По-перше, можна зберігати обидва значення, прив'язавши їх до різних шарів або каналів карти глибини. Фактично,

формується багатошарова карта глибини, де базовий шар – це звичайна глибина сцени, а додатковий шар у відповідних пікселях містить глибини відбитих та прихованих об'єктів. По-друге, можна реконструювати 3D-моделі сцени у вигляді мешу. З першої карти генерується меш, що моделює всі непрозорі об'єкти (включно з відбитими), а з другої карти – меш для самих відбивних поверхонь [20]. При цьому зберігається семантична інформація про об'єкти сцени. Об'єднуючи ці меші разом, отримуємо повну 3D-реконструкцію сцени, що включає в коректному положенні і відбиті об'єкти, і об'єкти приховані за напівпрозорими поверхнями. Прикладом результату може бути ситуація, коли на зображенні є дзеркало, що відбиває кімнату: традиційна одношарова глибина покаже діру або стіну замість дзеркала, або змішає відбиття із реальною сценою, тоді як запропонований метод дасть два значення – відстань до дзеркала і відстань до відбитих меблів, дозволяючи відобразити і те, і інше у 3D.

Двоканальна карта глибини з маскою використання каналів:

$$D(u) = (D_{surf}(u), D_{obj}(u)) = (\alpha_{surf}(u), \alpha_{obj}(u)), \quad (3.5)$$

де: $\alpha_{surf}(u) = p_{opaque}(u) + p_{translucent}(u) + p_{reflective}(u)$, $\alpha_{obj}(u) = p_{translucent}(u) + p_{reflective}(u)$.

Двозначні (двоканальні) карти глибини суттєво розширюють охоплення складних сцен. Варіанти реалізації двозначної карти глибини:

1) Сцена як карта відстаней та маска типу. Подати сцену як карту відстаней до непрозорих об'єктів, маску прозорості/відбивання та відстань до однієї (або жодної) прозорої/відбивної поверхні. Проста форма подання, що суттєво розширює охоплення складніших сцен і можливостей сценаріїв ДР.

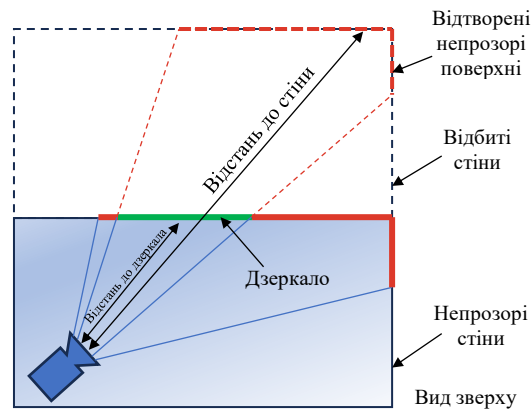


Рис. 3.5. Ілюстрація реалізації двозначної карти глибини за допомогою карти відстаней та маски типу.

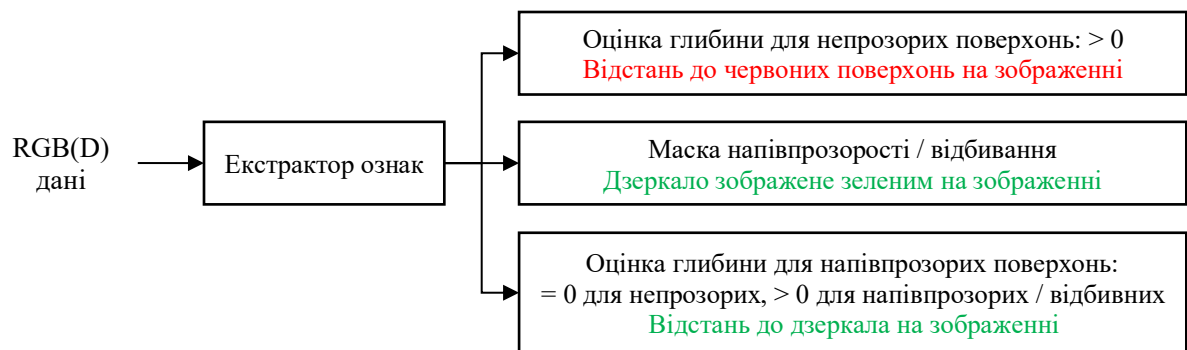


Рис. 3.6. Схематичне зображення реалізації двозначної карти глибини за допомогою карти відстаней та маски типу.

2) Сцена як набір фронтально-паралельних площин. Сцена подається як набір фронтально-паралельних площин, де кожне значення означає імовірність заповнення у відповідній точці. Щоб врахувати нефронтально-паралельні площини, використовується об'єм зсувів глибини. Значення цього об'єму задають величину зсуву для відповідної точки на фронтально-паралельній площині, щоб перемістити її на реальну поверхню об'єкта. Об'єм напівпрозорості/відбивності призначає коефіцієнти напівпрозорості та відбивності відповідним елементам об'єму імовірностей глибини.

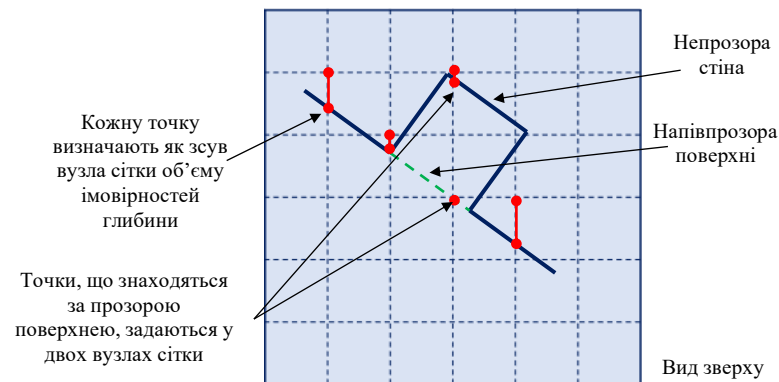


Рис. 3.7. Ілюстрація реалізації двозначної карти глибини за допомогою набору фронтально-паралельних площин.

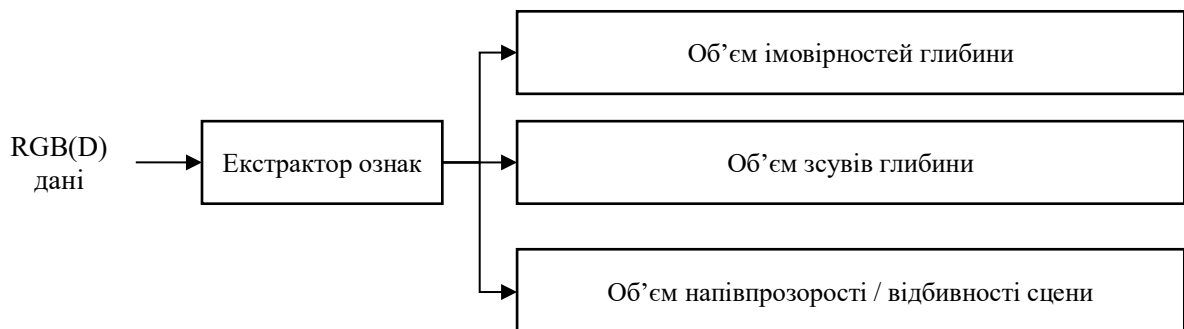


Рис. 3.8. Схематичне зображення реалізації двозначної карти глибини за допомогою набору фронтально-паралельних площин.

Варто зазначити, що реалізація такого алгоритму потребує надійних модулів комп'ютерного зору: сегментації об'єктів, оцінки глибини за допомогою нейронних мереж, комбінування даних різних сенсорів. Сучасний рівень технологій глибокого навчання дозволяє втілити ці компоненти. Існують нейронні мережі, що можуть детектувати дзеркала на зображенні, оцінювати відстань до них (наприклад, через аналіз відбитих об'єктів), а також високоточні моделі монокулярної глибини, які можна перенавчити з урахуванням складних випадків. Таким чином, запропонований метод є комплексним рішенням, що комбінує класичні підходи (стереозір), алгоритми глибокого навчання (оцінка глибини, сегментація) та концепцію багатопланового подання глибини для забезпечення повноти 3D-реконструкції.

3.3. Порівняння з існуючими підходами

3.3.1. Відмінності від традиційного стерео-відновлення глибини

У звичайному стереозорі (наприклад, камери в смартфонах з двома об'єктивами, чи стереокамери в робототехніці та окулярах ДР) алгоритм оцінює диспаратність шляхом порівняння двох зображень. Складні випадки, як ми розглянули, призводять до виникнення артефактів карт глибини. Області відбивних поверхонь можуть містити випадковий шум глибини або великі похибки, бо пікселі в лівому і правому зображеннях не відповідають одному реальному об'єкту. Ряд досліджень частково вирішували проблему за допомогою поляризаційних фільтрів, спеціальних шаблонів підсвічування або навіть аналізу багатократних відбиттів лазерних променів [115, 118]. Однак ці рішення ускладнюють систему апаратно або потребують спеціального налаштування. Запропонований метод та запатентований в [20], навпаки, працює на рівні програмної обробки даних і не вимагає нового апаратного забезпечення. Він може використовувати стереопару, але більш гнучко. За наявності кількох зображень сцени метод може оцінити глибину відбитого об'єкта, навіть якщо прямої відповідності немає, шляхом оптимізації під інші видимі частини об'єкта. Модулі оцінки карт глибини на нейронних мережах дозволяють методу вчитися на подібних випадках, чого не роблять класичні стерео підходи. Тому запропонований метод можна розглядати як надбудову над стереозором, що адресує специфічну проблему. Метод не відкидає дзеркальні області як нерелевантні, а обробляє їх окремо. Це якісно підвищує повноту карти глибини. Жодна область зображення не залишається без інтерпретації.

3.3.2. Порівняння з монокулярними мережами реконструкції карт глибини

Сучасні нейронні мережі для оцінки глибини (MonoDepth, MiDaS, DPT та інші) досягають вражаючих результатів на стандартних наборах даних. Вони спроможні відновлювати геометрію сцени з одного зображення, але мають

тенденцію помилятися на нетипових областях. Відбивні та напівпрозорі об'єкти якраз є нетиповими. Більшість датасетів для навчання (NYU-Depth, KITTI, MegaDepth та ін.) містять мало прикладів з дзеркальними поверхнями або взагалі їх виключають. Тому навіть найкращі моделі не дають правильної глибини на таких областях. Зазвичай вони або прогнозують велику глибину, фактично плутаючи відбиття з далеким фоном, або згладжують значення і дають некоректну середню глибину. В роботах відзначається, що CNN моделі добре захоплюють локальні деталі, але їм не вистачає глобального контексту, а трансформери навпаки бачать загальну картину, та можуть втрачати дрібні деталі [117]. У випадку відбивних поверхонь це проявляється так – CNN може помітити рамку дзеркала і припустити, що там порожнина (тому поставить далеко фон), а трансформер може зрозуміти, що це відбиття, але не знати, як оцінити глибину (оскільки ця глибина двозначна). Запропонований метод [20] кардинально різниться від таких підходів тим, що розв'язує неоднозначність через явне розділення двох шарів сцени. Можна сказати, що він накладає на стандартну модель глибини інтелектуальний постпроцесинг: якщо модель не знає, що робити з дзеркалом, алгоритм все одно витягне з неї користь – визначить, що за дзеркалом є об'єкт, і спробує визначити його глибину, а потім окремо додасть дзеркало. Цей підхід можна сумісно використовувати з будь-якою архітектурою нейромережі. У рамках методу можливо задіяти дві різні нейронні моделі: одну – для прогнозування базової глибини сцени, іншу – для оцінки глибини відбивних поверхонь. Таким чином, метод не конкурує безпосередньо з існуючими, а доповнює їх, пропонуючи системне рішення спеціального випадку. Варто відзначити, що вже з'являються дослідження, які адресують цю проблему навчанням моделей, стійких до відбивань [115], але вони здебільшого концентруються на покращенні одного значення глибини (напр., штрафуючи похибки на відблисках). На відміну від них, запропонований та запатентований метод [20] пропонує структурне покращення формату виходу у вигляді двошарового представлення глибини. Це більш універсальний підхід, який не залежить від того, наскільки добре модель навчилася оцінювати глибину

відбивної поверхні, адже система після отримання прогнозів сама коригує і розподіляє значення по своїх місцях (поверхня, або за поверхнею).

3.3.3. Використання методу в сучасному контексті

Розглянемо, як метод вписується в сучасний контекст. З одного боку, існують апаратні рішення проблеми – наприклад, LiDAR-сенсори можуть частково сприймати скляні стіни, але навіть вони не завжди дають інформацію про те, що за склом. Лазерний промінь може відбитися, не пройшовши крізь скло. У автонавігації використовують радарні сенсори, які дозволяють сканувати крізь скло, але мають низьку роздільну здатність. Таким чином, суто сенсорні методи повністю не вирішують проблему оцінки карт глибини при наявних відбивних чи напівпрозорих поверхнях. З іншого боку, програмні методи можуть працювати, використовуючи дані звичайних камер. Запропонований та запатентований метод [20] є однією з перших комплексних спроб розв’язати задачу “дзеркальної глибини” на рівні алгоритму побудови карти. Його конкурентами можна вважати наукові прототипи, наприклад, підхід, де використовують структуроване освітлення і поляризацію для вимірювання і глибин віддзеркалень, і об’єктів за ними [119, 120]. Проте такі системи поки що далекі від інтеграції у споживчі пристрої. Натомість, описуваний метод орієнтований на реалізацію у рамках програмного забезпечення, з використанням вже наявних апаратних компонентів та існуючих алгоритмів.

Отже, у порівнянні з альтернативами, метод вигідно вирізняється своєю здатністю доповнити стандартну карту глибини додатковою інформацією про складні області. Він інтегрує елементи і традиційного (стерео), і сучасних (CNN/Transformer) підходів, усуваючи їхні слабкі місця в окремо взятому сценарії. З огляду на тенденцію до використання гібридних моделей (CNN+Transformer) для підвищення точності глибини [117], можна очікувати, що ідеї запропонованого методу також можуть бути реалізовані у вигляді гібридної архітектури – наприклад, мережа, яка має дві “голови” для прогнозування двох карт глибини, та спільний модуль сегментації дзеркальних областей.

3.4. Висновки до розділу

Розглянутий метод генерації карти глибини з урахуванням відбивних і напівпрозорих об'єктів представляє собою значний крок у напрямі повноцінної 3D-реконструкції реальних сцен. Запропоноване рішення усуває ключові обмеження існуючих алгоритмів глибинного бачення, дозволяючи коректно моделювати сцени, що містять дзеркала, скляні перегородки та інші складні поверхні. Впровадження такого підходу може мати вагомий вплив на практичні застосування:

- Комп'ютерний зір та 3D-реконструкція. У задачах сканування приміщень та побудови 3D-моделей середовища наявність дзеркальних стін або вікон більше не буде призводити до прогалин у моделях. Система зможе відобразити як самі такі об'єкти, так і що за ними, підвищуючи повноту моделі. Це особливо цінно для окулярів ДР: приміщення з дзеркалами будуть відтворені правильно, і віртуальні об'єкти можуть взаємодіяти з ними реалістично (напр., відображатися у дзеркалі).
- Робототехніка та автономна навігація. Роботи й автономні транспортні засоби часто використовують камери та лідари для сприйняття перешкод. Дзеркальні поверхні можуть вводити їх в оману – відбиття виглядає як продовження простору, а скляна стіна може бути невидимою. Інтеграція методу, що розпізнає такі об'єкти і визначає їхню глибину, зробить навігацію безпечнішою. Наприклад, робот здатен завчасно виявити скляні двері як перешкоду, навіть якщо камера бачить крізь них далекий коридор. Автомобільний автопілот зможе відрізнити відблиск на дорозі від реальної калюжі або ями, знаючи, що це відображення не потребує маневру.
- Доповнена та віртуальна реальність. У доповненій реальності відображення реального світу у дзеркалах може спричиняти проблеми узгодження. Якщо система отримує від камер багаточисельну глибину,

вона може правильно рендерити віртуальні об'єкти. Наприклад, персонаж ДР буде видно у дзеркалі тільки тоді, коли він дійсно перед ним, і з правильним масштабом, оскільки відома геометрія дзеркала.

- Системи безпеки та спостереження. Алгоритми відеоаналітики зможуть краще оцінювати ситуацію в приміщеннях з великою кількістю відбивних поверхонь (наприклад, у магазинах з вітринами). Глибина допомагає виявляти об'єкти та їх переміщення, і якщо вона коректна навіть на віддзеркаленнях, зменшиться кількість хибних тривог (коли система “бачить” людину за склом двічі – реально і в відображенні – і сприймає це як двох різних людей).

Таким чином, розроблений метод має потенціал підвищити надійність та реалістичність комп'ютерного сприйняття простору у найрізноманітніших прикладних сценаріях. Він демонструє, як поєднання глибинного навчання та геометричних міркувань дозволяє покращити якість 3D реконструкції сцени при наявності на ній відбивних чи напівпрозорих поверхонь.

РОЗДІЛ 4: ПРОГНОЗУВАННЯ КАРТ ГЛИБИНИ ВИСОКОЇ ТОЧНОСТІ НА ПРИСТРОЯХ З ОБМЕЖЕНОЮ РОЗРЯДНІСТЮ ЗА ДОПОМОГОЮ ДВОВИМІРНИХ КРИВИХ ГІЛЬБЕРТА

4.1. Проблематика зниження якості реконструкції карт глибини та їх діапазону при портуванні методів на DSP/NPU

Останніми роками прогнозування карт щільної глибини привертає значну увагу завдяки широкому спектру застосувань у таких галузях, як розуміння середовища [113, 121], автономне водіння [122, 123], робототехніка [124], доповнена та віртуальна реальність [125], інтернет речей [126], безпілотники [127, 128]. Як правило, застосування у цих сферах обмежене ресурсами обчислювальної потужності та енергоспоживанням. Точна оцінка карти глибини відіграє ключову роль у забезпеченні надійного сприйняття середовища та прийняття рішень. Більшість сучасних рішень для монокулярної [78, 129–134] та бінокулярної [135, 136] оцінки карти глибини базуються на великих фреймворках з використанням складних згорткових нейронних мереж (включно з Convolutional Gated Recurrent Units) [135–137], трансформерів [129–131] і дифузійних архітектур [78, 132–134]. Ця тенденція пов'язана з прагненням розробити фундаментальні моделі для розв'язання задачі. Проте такий прогрес супроводжується зростанням складності моделей, що вимагає значних обчислювальних ресурсів та обсягів пам'яті. Обмеження апаратного забезпечення створюють додаткові труднощі при розгортанні таких моделей на кінцевих пристроях.

Підвищення ефективності виконання DNN моделей на малопотужних пристроях можливе за рахунок кількох підходів. Один із них – оптимізація архітектури моделі шляхом зменшення кількості параметрів або затримки, уникнення використання обчислювально складних шарів, використання прорідження мережі [124, 138–140]. Інший підхід – використання обчислень в низькій розрядності (точності) [141, 142]. Представлення ваг і активацій DNN моделі у вигляді фіксованої розрядності, наприклад INT8, дозволяє збільшити

паралельність обчислень, зменшити обсяг передачі даних і скоротити енергоспоживання операцій Multiply-Accumulate (MAC) [143]. Зменшення розрядності веде до лінійного зменшення обсягу необхідної пам'яті й квадратичного зниження складності матричних множень [25]. Моделі високої точності зазвичай навчають у форматі FP32 і квантують до формату з низькою розрядністю для виконання на кінцевих пристроях. Квантування з урахуванням навчання (QAT) емулює процес квантування під час навчання моделі [144–147]; квантування після тренування (PTQ) працює з уже навченими моделями та намагається зменшити втрати якості при переводі ваг у цілочисельний формат [147, 148].

Проте виконання квантованих моделей призводить до появи додаткових помилок обчислень і втрати точності [143]. Багато досліджень спрямовано на зменшення похибки квантування шляхом модифікації архітектури DNN [149, 150], покращення самих технік квантування [143, 145–147, 147, 148], адаптації до специфіки апаратного забезпечення [151–153]. Проблема втрати точності даних майже не досліджується, імовірно через те, що вона не критична для моделей, які прогнозують RGB-зображення, природно представлені трьома 8-бітними каналами. Однак для зображень із високим динамічним діапазоном втрата точності стає критичною. Це обмежує застосування ефективного однорідного 8-бітного квантування і змушує використовувати змішані схеми квантування [154].

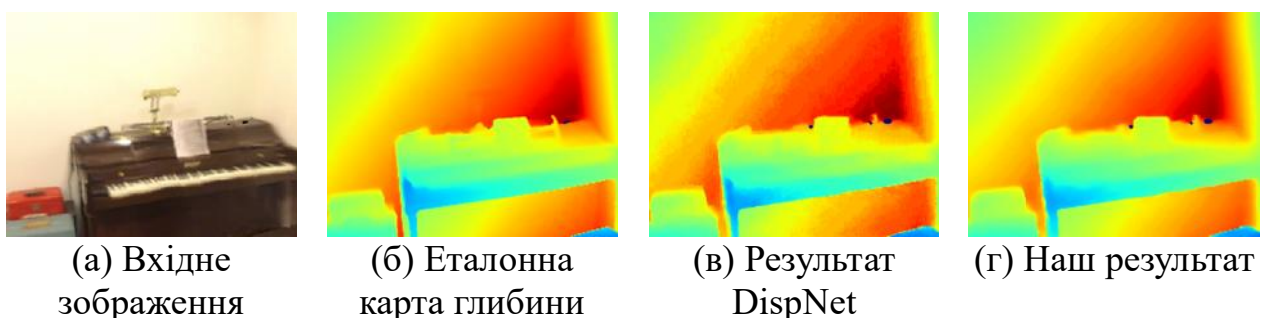


Рис. 4.1. Ілюстрація квантування моделі DispNet [94] з INT8 точністю (W8A8) [21]. Запуск квантованої моделі на Qualcomm Hexagon DSP-процесорі призводить до втрати точності карти глибини та появи квантувальних артефактів (в). Запропонований нами підхід дозволяє збільшити розрядність карти глибини та зменшити похибку квантування (г).

Карти глибини вимагають високої розрядності для точного представлення. Наприклад, для представлення глибини в діапазоні від 0 до 10 м з точністю 1 см необхідно 10 біт. Восьмибітне представлення карти глибини призводить до появи помилкових меж на площинах (Рис. 4.1, в), втрати просторових деталей та спотворення об'єктів з низьким контрастом глибини. Як наслідок, малопотужні пристрої з низькорозрядною арифметикою неминуче втрачають якість прогнозування карти глибини. У [101] для подолання цього обмеження запропоновано залишити у високій точності останній шар моделі, а решту – квантованими до 4 або 8 біт. Проте це рішення не є ефективним, оскільки останній шар обробляє дані в найвищій просторовій роздільній здатності, що потребує значних ресурсів. Не всі пристрої підтримують змішане квантування або різні розрядності для різних шарів. Наприклад, бібліотека SNPE [155], яка використовується для виконання моделей на Qualcomm Hexagon DSP, вимагає однакову розрядність для всіх шарів (окремо для ваг і активацій). Апаратний прискорювач для обрахунку тензорів Coral Edge TPU підтримує лише INT8 або UINT8, усі операції з числами з плаваючою точкою виконуються на CPU [156]. Отже, на пристроях з низькорозрядною арифметикою моделі для прогнозування глибини неминуче страждають від втрати точності незалежно від якості квантування. Це обмеження зумовлене апаратною платформою й не може бути усунуте покращенням QAT та PTQ підходів.

У цьому розділі запропоновано подолати це обмеження за допомогою представлення глибини як точок на 2D кривій Гільберта [21]. Така трансформація кодує глибину з високим динамічним діапазоном як дві компоненти кривої Гільберта, що можуть бути представлені у форматі з низькою розрядністю та мінімальними втратами по якості (Рис. 4.1). Запропонований підхід передбачає додатковий етап постобробки на пристрої, що відновлює точну карту глибини з квантованих компонент кривої Гільберта. Постобробка не вимагає операцій додавання чи множення та може бути реалізована у вигляді таблиці підстановки (LUT). Основна відмінність запропонованого підходу – для підвищення точності прогнозування карт глибини враховується не лише архітектура DNN моделі,

процес навчання (наприклад QAT), ефективне квантування ваг та активацій (PTQ), а й структуру сигналу, який прогнозується моделлю. Цей підхід не змінює сам процес квантування, а використовується разом з існуючими методами (PTQ або QAT), тож прогрес у їх розвитку також підвищує ефективність запропонованого підходу.

Основні задачі розділу можна узагальнити наступним чином:

1. Запропонувати новий підхід до якісного прогнозування глибини широкого діапазону на пристроях з низькорозрядною арифметикою.
2. Оцінити ефективність методу і показати, що модифікована квантована модель може досягти подібної або кращої якості прогнозування, ніж оригінальна модель.
3. Оцінити вплив запропонованого методу на якість та похибку квантування моделі.

4.2. Метод

У цьому розділі представлено метод, що дозволяє здійснювати прогнозування карт глибини високої точності на пристроях з обчисленнями в низькій розрядності [21]. Основна ідея методу полягає у представленні глибини як двох компонентів двовимірної кривої Гільберта. Безпосередньо на пристрої ці компоненти прогнозуються в низькорозрядному форматі на TPU, NPU або DSP, а потім перетворюються у високорозрядне значення карти глибини за допомогою простого алгоритму постобробки, що виконується на CPU.

4.2.1. Прогнозування високоточних карт глибини на пристроях з низькою розрядністю

Розглянемо DNN модель, яка прогнозує деяку величину q , обмежену діапазоном $[0,1]$. Квантована модель виконується на кінцевому пристрої, що має апаратний прискорювач для виконання DNN (у нашому випадку це DSP) з b -бітним представленням вихідних даних та CPU загального призначення, що

проводить обчислення в повній точності. DSP повертає значення $q_{quant.b}$ з точністю b -біт (наприклад, INT8). Різниця між q та $q_{quant.b}$ називається похибкою квантування. Позначимо стандартне відхилення (SD) цієї похибки як σ_{quant} .

Метою є збільшення розрядності прогнозованого значення q при мінімальних додаткових обчислювальних витратах [21]. У випадку прогнозування карт глибини величина q відповідає нормалізованому значенню карти глибини або нормалізованій бінокулярній диспаратності, залежно від архітектури DNN.

Через обмеження DSP розрядність вихідних даних моделі не може бути збільшена, тому ми можемо оперувати лише кількістю та структурою вихідних каналів. Ці канали мають бути передані з DSP до CPU для відновлення q у форматі з більшою розрядністю. Щоб така схема була ефективною, потрібно вирішити дві ключові задачі:

1. мінімізувати обсяг даних, що передаються з DSP на CPU;
2. мінімізувати складність постобробки на CPU.

У рішенні, запропонованому в [157], останній шар моделі для відтворення глибини залишено у форматі з плаваючою точкою. Це рішення є неефективним, оскільки потребує передачі великого обсягу даних з передостаннього шару до CPU і обробки карти у високій роздільній здатності в останньому шарі.

Ми пропонуємо представляти q як точку на 2D параметричній кривій $(x(q), y(q))$ з довжиною $L > 1$, де обидві координати $x(q)$ та $y(q)$ обмежені інтервалом $[0,1]$. Модель з повною точністю навчається безпосередньо прогнозувати $x(q)$ та $y(q)$. Під час виконання на DSP значення q обчислюється з x та y значень, які передбачені в b -бітному форматі. Крива довжиною L проходить через приблизно $L \cdot 2^b$ дискретних точок $(x(q), y(q))$, що ефективно підвищує точність реконструкції q на $\log_2 L$ біт з b до $b + \log_2 L$ (див. Рис. 4.2). У такій реалізації обсяг даних для передачі з DSP до CPU зростає лише вдвічі, постобробка є простою і реалізується через LUT [21]. Надалі немодифікована

DNN називається «оригінальною моделлю», а модель, яка прогнозує компоненти кривої Гільберта (разом з постобробкою) – «модифікованою моделлю».

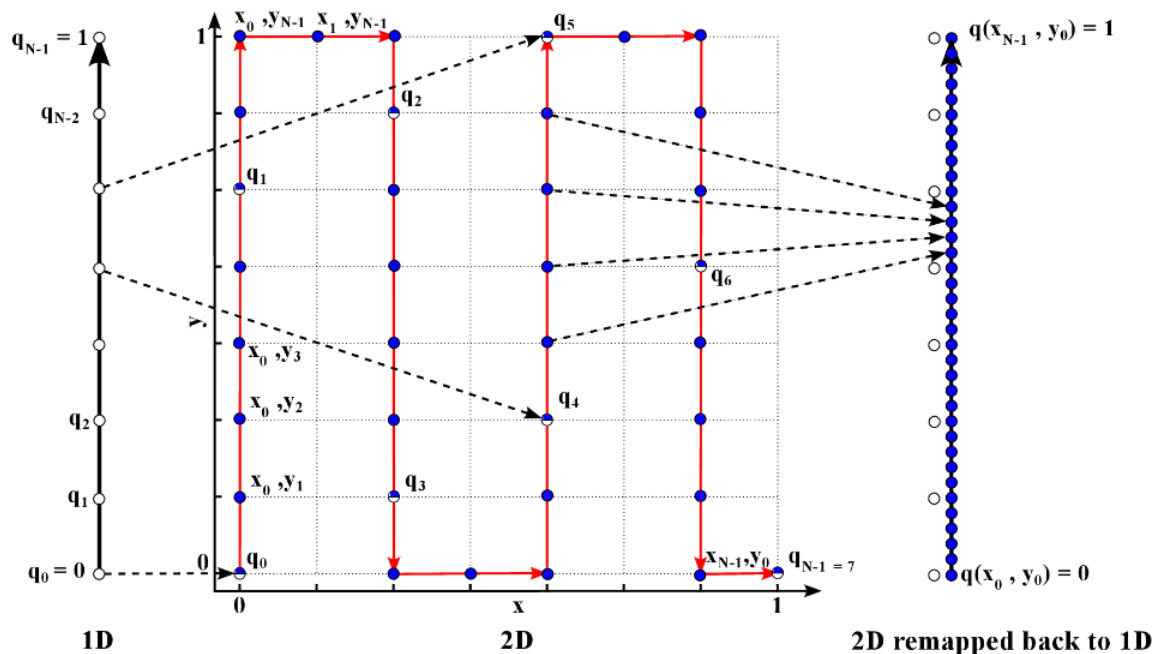


Рис. 4.2. Ілюстрація ідеї [21]. Одновимірний діапазон квантується до $N = 8$. Значення $q_0 = 0 \dots q_{N-1} = 1$ позначені білими кружечками. Цей одновимірний діапазон відображається на двовимірну криву, показану червоним кольором. Вісі x та y також квантуються до $N = 8$ значень, утворюючи 64 двовимірні точки. Серед них 36 точок лежать на кривій (позначені синім кольором). Відображення двовимірної кривої назад в одновимірний діапазон дає 36 різних квантованих значень. Таким чином, похибка квантування ефективно зменшується в $L = 35 / 7 = 5$ разів.

Розгортання DNN моделі для прогнозування карт глибини на кінцевому пристрої складається з наступних етапів (Рис. 4.3):

1. навчання моделі з повною точністю для прогнозування компонент кривої Гільберта, що представляють карту глибини;
2. застосування стандартного методу квантування (PTQ або QAT);
3. виконання модифікованої квантованої моделі на пристрої та отримання компонентів кривої Гільберта у низькій розрядності;
4. постобробка компонентів кривої Гільберта та відновлення глибини з високою розрядністю.

Вибір функції $(x(q), y(q))$ є критичним для успішної реалізації запропонованого підходу.

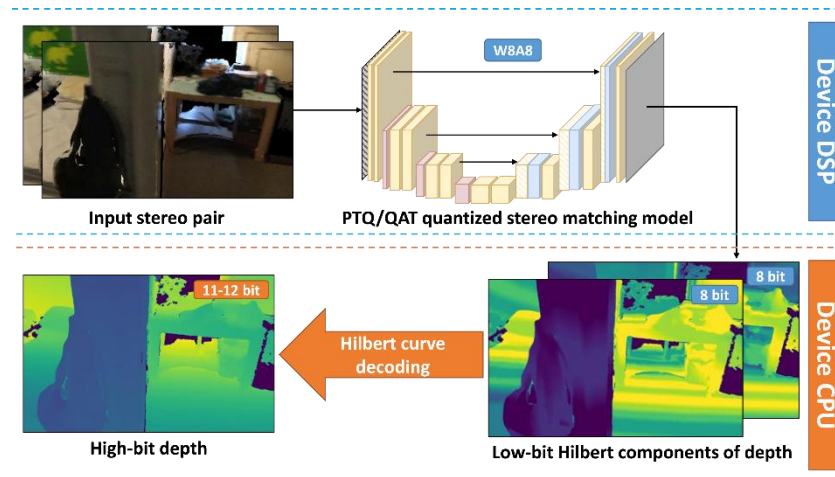


Рис. 4.3. Схема виконання запропонованого методу на пристрої [21].

4.2.2. Вибір оптимальної параметричної кривої

Розглянемо найпростішу функцію: $x(q) = \lfloor q * 255 \rfloor / 255$, $y(q) = q - x(q)$, де $\lfloor \cdot \rfloor$ – операція округлення до меншого значення. У цьому випадку $x(q)$ є грубим наближенням q , а $y(q)$ додає деталізацію (див. Рис. 4.4, д та Рис. 4.4, е). Довжина цієї кривої становить 256 і дозволяє реконструювати q з точністю INT16 при $x(q)$ і $y(q)$ з точністю INT8. Однак структура $y(q)$ є несприятливою для навчання моделі з повною точністю: вона має багато різких переходів між нулем і одиницею. Під час експериментів було встановлено, що така крива погіршує навчання моделі з повною точністю і дуже вразлива до помилок квантування.

Зазначимо бажані властивості кривої $(x(q), y(q))$:

- Неперервність: незначні зміни q повинні призводити до незначних змін $x(q)$ та $y(q)$. Це гарантує, що $x(q)$ та $y(q)$ зберігають просторову плавність та не вводять додаткових розривів значень при кодуванні глибини.
- Однозначність: крива не повинна самоперетинатися, тобто має забезпечувати взаємно однозначну відповідність між q та $(x(q), y(q))$.
- Рівномірне покриття: крива має рівномірно покривати одиничний квадрат і уникати граничних точок $(x(q_1), y(q_1))$ та $(x(q_2), y(q_2))$ при віддалених q_1 та q_2 .

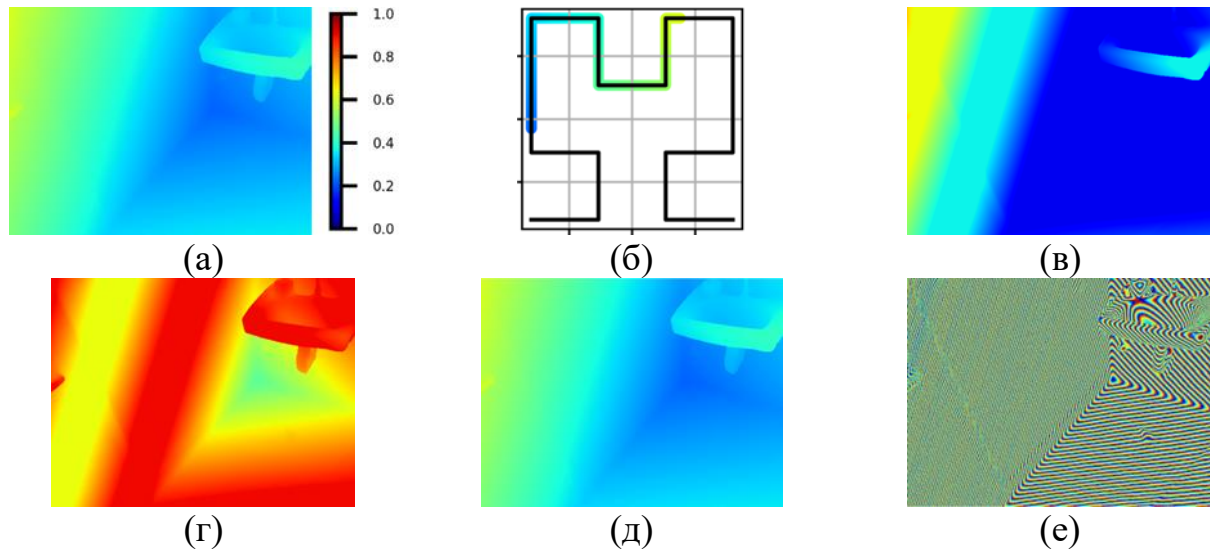


Рис. 4.4. Ілюстрація перетворень диспаратності [21]: (а) карта диспаратності; (б) відображення в 2D-простір за допомогою кривої Гільберта другого порядку; (в, г) компоненти x та y кривої Гільберта; (д, е) грубі та точні деталі карти диспаратності. Точні деталі на (е) відповідають найменш значущому байту диспаратності (а), представленого у 16-бітному форматі. Високочастотні коливання створюють візуально відмінне зображення від оригінальної карти диспаратності, що ускладнює їх передбачення моделлю глибокого навчання.

Криві з такими властивостями відомі як криві, що заповнюють простір (space-filling curves), або криві Пеано [158]. Для запропонованого методу найбільш простою та гнучкою виявилася саме крива Гільберта [19, 158].

Крива Гільберта є представником широкого класу кривих заповнення простору (space-filling curves) [159]. Далі представлені додаткові аргументи на користь вибору саме кривої Гільберта для задач прогнозування карт глибини з високою точністю. У цьому аналізі витримана термінологія з роботи [159], де криві класифікуються з використанням квадратичної і трикутної сітки та поділяються на родини \sqrt{N} . У кожній родині \sqrt{N} відстань між початковою та кінцевою точками генератора кривої дорівнює \sqrt{N} .

Для повноцінного використання представлення карти глибини у вигляді двох компонент бажано, щоб крива заповнення простору рівномірно покривала одиничний квадрат. Це вимога одразу виключає криві, побудовані на трикутних сітках.

Криві з неортогональними генераторами (наприклад, крива Z-порядку) мають недолік – вони нерівномірно заповнюють одиничний квадрат. Серед кривих, оснований на квадратній сітці та використовуючих ортогональну генерацію, вибір обмежується кривою Гільберта (родина $\sqrt{4}$, Рис. 4.5, а), кривою Пеано (родина $\sqrt{9}$, Рис. 4.5, б) та квадратичною кривою Госпера (родина $\sqrt{25}$, Рис. 4.5, в). Для кривих Гільберта та Пеано можливі різні генератори. Наприклад, крива Мура є варіантом кривої Гільберта. Вони відрізняються лише порядком заповнення простору.

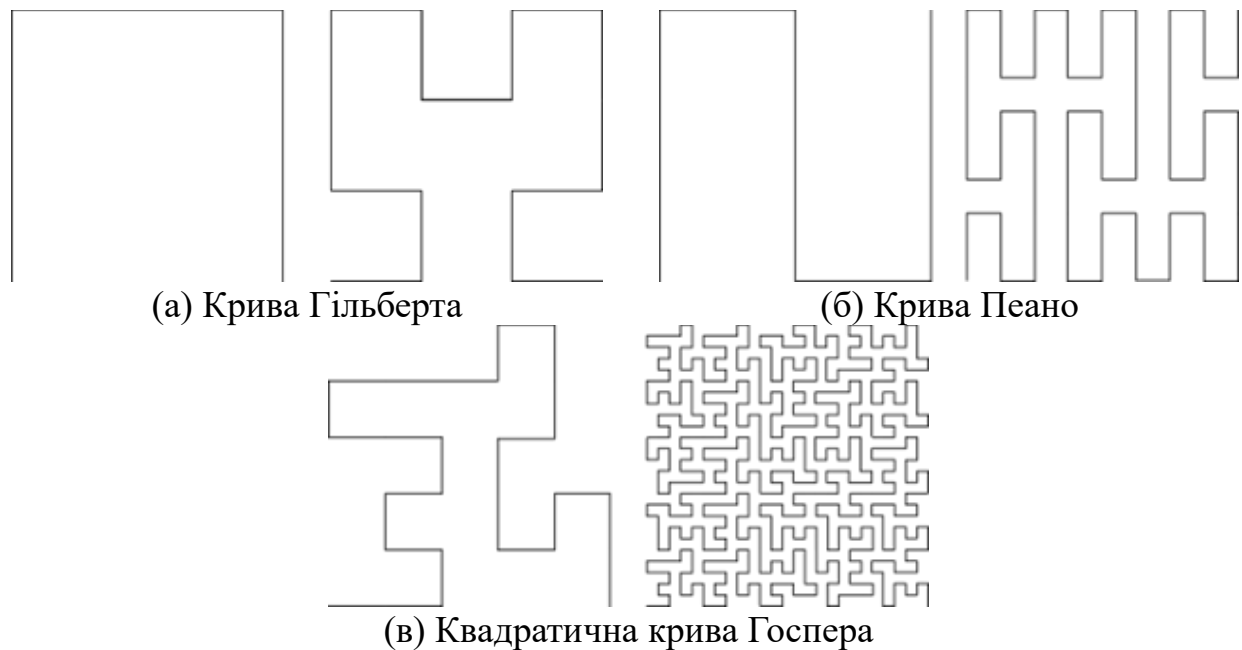


Рис. 4.5. Криві, що заповнюють простір, у межах одиничного квадрата [19]. Для кожної пари продемонстровані криві першого (зліва) та другого (справа) порядку p .

Значення N визначає, з якою швидкістю зростає довжина кривої L_p та зменшується розмір її відрізка h_p зі збільшенням порядку кривої. Експерименти показали, що бажано мати можливість підлаштовувати довжину кривої залежно від величини похибки квантування моделі. З цієї точки зору, крива Гільберта є найгнучкішою оскільки має найменше значення N . Для неї, зі збільшенням порядку p , кількість вузлів зростає наступним чином: 1, 4, 16, 64, 256. Для кривої Пеано – 1, 9, 81, 729, 6561, а для квадратичної кривої Госпера – 1, 25, 625, 15625, 390625 [21]. Якщо обмежити кількість вузлів до розумного значення 256, то крива

Гільберта забезпечує 4 придатні криві нижчого порядку (які були використані в експериментах), Пеано – 2, квадратична крива Госпера – 1.

Крива Гільберта є найпростішою та найгнучкішою серед тих, що відповідають усім необхідним вимогам для кодування значень глибини. Для побудови ще гнучкішої послідовності кривих можна використати криві Гільберта, Пеано та квадратичну криву Госпера різних порядків для створення набору кривих з кількістю вузлів: 4, 9, 16, 25, 64, 81, 256.

Якщо дотримано основні вимоги до параметричної кривої (відсутність самоперетинів, рівномірне заповнення одиничного квадрата, неперервність), то детальна структура кривої не має принципового значення. Наприклад, можна використовувати довільні криві, що заповнюють квадрат із заданою кількістю вузлів, криві з округленими кутами або криві, які по-різному стискають ділянки 1D-величини (щоб виділити діапазони з найбільш імовірними значеннями цільової величини).

Крива Гільберта – це неперервна фрактальна крива, що заповнює простір, побудована як обмежена послідовності кусково-лінійних кривих [19]. Вона починається з однієї точки в центрі одиничного квадрата. Кожен наступний порядок кривої утворюється шляхом відтворення та з'єднання точок кривої попереднього порядку. Позначаємо порядок кривої як p . Полігони, що апроксимують криві порядку 1–4, показано на Рис. 4.6. Щоб уникнути крайових ефектів, крива маштабована так, щоб вона вміщалася в квадрат $\{(x, y) | b \leq x, y \leq 1 - b\}$, де $b = 0.1$. У цьому випадку довжина кривої p -го порядку становить $L_p = (2^p + 1)(1 - 2b)$. Довжина границі полігона, що описує криву дорівнює $h_p = (1 - 2b) / (2^p - 1)$ – ця величина також задає мінімальну відстань між точками різних паралельних границь полігону, що обмежує криву Гільберта.

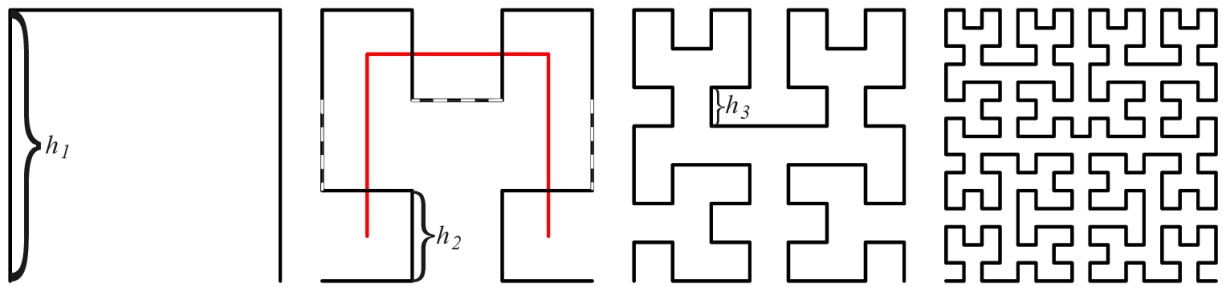


Рис. 4.6. Криві Гільберта для порядків $p = 1, 2, 3, 4$ (зліва направо) [21]. Кожен наступний порядок формується шляхом заміни кожної вершини на послідовність із трьох сегментів.

4.2.3. Пряме та зворотне перетворення

Для моделі з повною точністю та квантованої моделі передбачені значення (x, y) не мають точної відповідності з будь-якими $(x(q), y(q))$. Щоб перетворити довільну точку (x, y) назад у одновимірне значення, знаходиться найближча до неї точка на кривій Гільберта [21]:

$$q_{xy} = \arg \min_{q \in [0,1]} \|x - x(q), y - y(q)\| \quad (4.1)$$

Побудова прямого ($1D \rightarrow 2D$) та зворотного ($2D \rightarrow 1D$) відображення для кривих Гільберта довільного порядку базується на ітеративних алгоритмах [19]. Оскільки використовується крива фіксованого порядку, можна реалізувати швидші перетворення за допомогою таблиць підстановки (LUT). Використані дві LUT для відповідних відображень. Перша LUT створюється шляхом білінійної інтерполяції вузлів кривої Гільберта низького порядку. Вона дозволяє отримувати $(x(q), y(q))$ для заданого q . На Рис. 4.4 (а, б) наведено приклади такого перетворення застосовані до еталонної (GT) карти диспаратності. Для відображення 2D значень (x, y) (Рис. 4.4 (в, г)) назад у 1D представлення використовується друга LUT, побудована за формулою (4.1). В наступних експериментах ця таблиця має розмір 256 на 256 елементів.

4.2.4. Перетворення похибки квантування

Для оригінальної моделі похибка квантування вимірюється безпосередньо на виході моделі. Для запропонованого підходу похибка оцінюється після

постобробки компонентів кривої Гільберта і включає в себе як похибку квантування самих компонентів, так і вплив трансформації під час постобробки [21]. Стандартне відхилення (SD) цієї похибки – $\bar{\sigma}_{quant}$, а SD похибки квантування компонентів кривої Гільберта – $\sigma_{xy.quant}$.

Локально, рівняння (4.1) зводиться до проекції точки (x, y) на найближчий відрізок кривої Гільберта з відкиданням однієї з компонент (x або y залежно від орієнтації відрізка), а також до стискання іншої компоненти в L разів. Як зазначено вище, ця трансформація додає $\log_2 L$ біт до точності q . Крім того, похибка квантування компонентів кривої Гільберта також стискається в L разів, що дає: $\bar{\sigma}_{quant} = \sigma_{xy.quant} / L$.

Стиснення похибки квантування можливе, якщо: (а) похибка квантування компонентів кривої Гільберта є незалежною між каналами та однаково розподіленою; (б) значення $\sigma_{xy.quant}$ є достатньо малим, щоб похибка квантування компонент кривої Гільберта була меншою за h_p . Оскільки оригінальна та модифікована моделі мають подібну архітектуру, вирішують ту ж задачу, навчені на однакових даних та квантовані тим самим методом, можна додатково припустити, що (в) $\sigma_{quant} = \sigma_{xy.quant}$. Якщо це припущення справедливе, то запропонований підхід дозволяє зменшити похибку квантування в L разів: $\bar{\sigma}_{quant} = \sigma_{quant} / L$.

Умова (б) забезпечується шляхом вибору доцільного порядку кривої Гільберта. Для стереозіставлення експериментально встановлено, що порядки $p = 2, 3$ є оптимальними. Вони дають довжину кривої 4 – 7.2 та збільшення розрядності на 2 – 2.85 бітів. Припущення (а) та (в) можна перевірити лише експериментально для конкретної архітектури моделі. Варто відмітити, що збільшення розрядності має місце навіть без виконання умов (а) та (в). Ці додаткові умови важливі лише для потенційного зменшення похибки квантування.

4.2.5. Модифікація моделі DispNet та відповідної функції втрат

Для реалізації запропонованого підходу необхідно змінити DNN модель з однією вихідною головою для передбачення q на модель з двома головами, які передбачають компоненти кривої Гільберта x та y . Цю модифікацію для моделі DispNet [94] показано на Рис. 4.7.

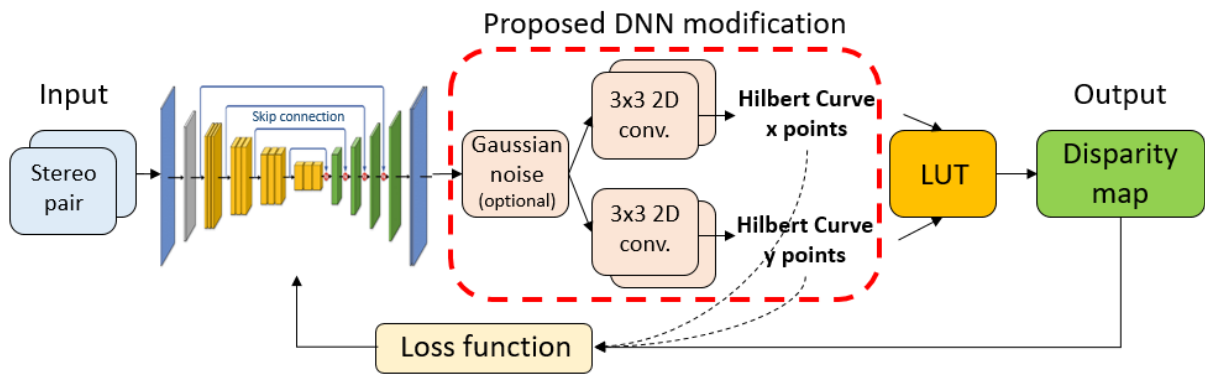


Рис. 4.7. Модифікація оригінальної моделі DispNet [21].

Модифікація моделі DispNet [94], необхідна для реалізації запропонованого підходу. Вхідна стереопара RGB обробляється енкодер-декодерною мережею з оригінальної моделі. Ознаки, отримані на виході енкодер-декодера, подаються на додатковий шар Гаусового шуму, після чого проходять через дві однакові голови для визначення компонент кривої Гільберта. На етапі постобробки ці компоненти перетворюються у фінальну карту диспаратності.

Функція втрат для запропонованого підходу складається з двох компонент: втрати за глибиною $\Lambda(q_{GT}, q_{xy})$ [160] та додаткової складової $\Lambda_H(x_{GT}, y_{GT}, x, y)$, яка забезпечує збіжність моделі до представлення на базі кривої Гільберта:

$$\Lambda_{full} = \Lambda(q_{GT}, q_{xy}) + \alpha \cdot \Lambda_H(x_{GT}, y_{GT}, x, y), \quad (4.2)$$

де x_{GT} та y_{GT} – це еталонні значення компонент кривої Гільберта, обчислені з еталонного значення q_{GT} , а α – гіперпараметр. Додатковий компонент Λ_H обчислюється за формулою:

$$\Lambda_H(x_{GT}, y_{GT}, x, y) = (x_{GT} - x)^2 + (y_{GT} - y)^2 + \beta \cdot r_{xy}^2, \quad (4.3)$$

де $r_{xy} = \|(x - x(q_{xy}), y - y(q_{xy}))\|$, а β – додатковий гіперпараметр. Компонент втрат по кривій Гільберта виконує дві функції: штрафує відстань між еталонними

значеннями (GT) та передбаченими точками у 2D вигляді; а також штрафує відхилення від кривої Гільберта, змушуючи модель передбачати лише точки, що належать до кривої. У експериментах були встановлені $\alpha = 1$ та $\beta = 25$.

4.2.6. Модифікація моделі DPT та відповідної функції втрат

Однією з моделей, вибраних для експериментів, є Dense Prediction Transformer (DPT) [129]. Усі архітектурні модифікації моделі проілюстровано на Рис. 4.8. Додатковий шар 2D згорткок 1×1 був доданий для коректної інтеграції вхідної стереопари RGB у ядро архітектури MobileViTv3-S [161].

Крім того, блоки MobileNet в енкодері були змінені для кращої сумісності з квантуванням згідно з підходом, описаним у [149]. Для прогнозування диспаратності використано оригінальну голову DPT. Архітектура голови для компонентів кривої Гільберта включає додатковий шар підвищення роздільної здатності після першої 3×3 2D згортки.

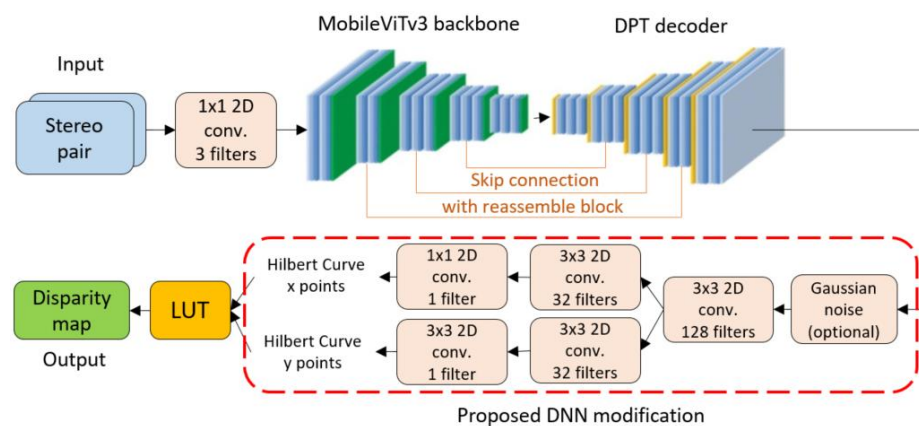


Рис. 4.8. Модифікація оригінальної моделі hrDPT [21].

Вхідна RGB-стереопара обробляється енкодером на основі MobileViTv3-S (виконує роль основної мережі) та декодером, запропонованим для задачі передбачення глибини в DPT. Ознаки, отримані з декодера, передаються на додатковий шар гаусового шуму та 3×3 2D-згортковий шар, після чого подаються на дві голови компонент кривої Гільберта. Кожна з них складається з одного 3×3 та одного 1×1 2D-згорткових шарів зі зменшенням кількості фільтрів: 32 та 1

відповідно. На етапі постобробки компоненти кривої Гільберта перетворюються у фінальну карту диспаратності.

Модель DPT містить енкодер MobileViTv3-S зі скіп-зв'язками перед кожним блоком MobileViT. Кожний скіп-зв'язок інтегрується у декодер за допомогою блоку повторної збірки (reassemble block), запропонованого в [129].

Під час аналізу архітектури мережі ми виявили, що шари в блоках MobileNet мають розподіли ваг із великим значенням коефіцієнту ексцесу. Як зазначено в [162], великі значення коефіцієнту ексцесу можуть призводити до погіршення якості квантування через обрізання викидів. Щоб зменшити похибки квантування в експериментах із прогнозуванням карт глибини та виходами у вигляді компонентів кривої Гільберта, була додана регуляризація коефіцієнту ексцесу, запропонована в [162], до всіх 1×1 згорток у блоках MobileNet моделі MobileViTv3-S.

4.3. Експерименти

Для експериментів зі стереозіставлення обрано дві моделі: DispNet з оригінальною архітектурою, запропонованою в [94], та Dense Prediction Transformer (DPT) [129] з MobileViTv3-S [161] в якості енкодера. У всіх експериментах розмірність входу моделі становить 384×512 пікселів, а виходу в -192×256 пікселів. Голова для прогнозування компонентів кривої Гільберта в моделі DPT така сама, як і для DispNet, однак містить додатковий згортковий шар та шар збільшення роздільної здатності в кожній гілці для адаптації до форм ознак декодера DPT. Виявлено, що додавання невеликої кількості гаусівського шуму на початку голови компонент Гільберта покращує квантування модифікованих моделей з використанням бібліотеки SNPE і не впливає на немодифіковані моделі. Експериментальні результати для модифікованих моделей представлені з шаром гаусівського шуму, де SD дорівнює 0.02.

4.3.1. Деталі реалізації

Для навчання моделей стереозіставлення було адаптовано ScanNet v2 датасет [113] наступним чином. Навчальні, валідаційні та тестові дані рендеруються з 3D реконструкцій, наданих для кожної сцени ScanNet v2, використовуючи бібліотеку PyRender v.0.1.45. Позиції камери для лівої камери фіксовані відповідно до позицій зазначених в наборі даних ScanNet v2. Права камера зміщена на 60 мм по осі x для формування горизонтальної бази стереопари. Внутрішні параметри лівої та правої камер відповідають даним ScanNet: пінхол камера з $f_x = f_y = 577.87$, $c_x = 320$, $c_y = 240$. Поділ на навчальну та тестову вибірки відповідає офіційному поділу набору даних ScanNet v2.

Усі моделі були квантовані за допомогою SNPE SDK v.2.24 та протестовані на пристрої Samsung S24+ із процесором Qualcomm Snapdragon 8 Gen3 та Hexagon DSP. Порівняно моделі у форматах FP16, W8A16 та W8A8, що запускаються на Hexagon DSP. Споживання енергії вимірювалось за допомогою Monsoon Solutions FTA22D Power Monitor у режимі енергозбереження.

4.3.2. Метрики оцінки якості

Якість передбачуваної глибини оцінюється за допомогою стандартних метрик [163, 164]: середня абсолютна відносна похибка (AbsRel), середньоквадратична похибка (RMSE), похибка кінцевої точки (EPE) та метрика D1.

Піксельні похибки самі по собі не повною мірою відображають артефакти квантування на картах глибини. Наприклад, представлення глибини у форматі INT8 майже не впливає на метрику AbsRel. Для вирішення цієї проблеми проведені експерименти з метрикою SSIM [165, 166]. Проте було виявлено, що вона майже не чутлива до артефактів квантування. Тому запропоновано використовувати косинусну подібність [167] між коефіцієнтами дискретного косинусного перетворення (DCT) [168] еталонних та передбачених карт глибини. Для цього $n \times n$ DCT за принципом скануючого вікна застосовується до обох карт

глибини – еталонної та передбаченої. Коефіцієнти DCT перетворюються у вектори, при цьому нульові коефіцієнти відкидаються. Косинусна подібність між отриманими векторами обчислюється в кожному положенні скануючого вікна, після чого усереднюється:

$$S_c = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M \frac{\mathbf{c}_{ij} \cdot \hat{\mathbf{c}}_{ij}}{\|\mathbf{c}_{ij}\| \cdot \|\hat{\mathbf{c}}_{ij}\|}, \quad (4.4)$$

де N – кількість кадрів, M – кількість вікон, \mathbf{c}_{ij} та $\hat{\mathbf{c}}_{ij}$ – вектори DCT коефіцієнтів для еталонних та передбачених карт глибини для кадру i та вікна j . Експерименти показують, що S_c , обчислена у вікні 4×4 , є чутливою до розмиття глибини та артефактів квантування INT8. Значення S_c , наближене до одиниці (максимально можлива подібність), свідчить про якісні карти глибини з чіткими краями та відсутністю артефактів в однорідних областях.

Для оцінки похибки квантування використано стандартне відхилення $\hat{\sigma}$ між виходами моделі з високою точністю та квантованої моделі, виміряне як масштабоване медіанне абсолютне відхилення (Scaled Median Absolute Deviation) [169].

4.3.3. Аналіз моделей у форматі W8A8

Були навчені моделі DispNet та DPT для порядків кривої Гільберта $p = 1, 2, 3, 4$. У подальшому ці моделі позначатимуться hpDispNet та hpDPT відповідно. Під час експериментів спостерігалось, що моделі високої точності hpDispNet та hpDPT навчаються передбачати точки, близькі до кривої Гільберта. Точність прогнозування карт глибини для модифікованих моделей є співставною з оригінальними моделями за всіма метриками.

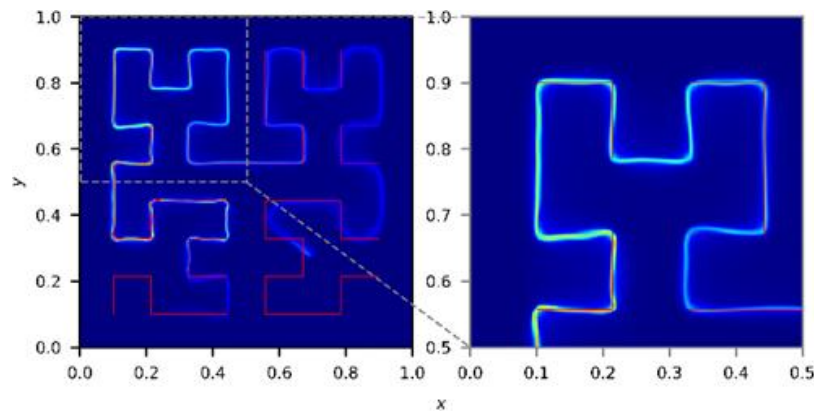
Було порівняно оригінальні та модифіковані моделі у форматі W8A8, які запускаються на CPU та DSP. Різниця полягає в тому, що на CPU модель W8A8 працює в режимі деквантування з використанням арифметики повної точності, тоді як на DSP виконання відбувається з обмеженою точністю. Кількісні результати наведено в Табл. 4.1. Для оригінальної моделі DispNet квантування

призводить до помітного погіршення якості як на CPU, так і на DSP. Зокрема, на CPU показник S_c знижується з 0.86 до 0.68, що вказує на втрату просторових деталей.

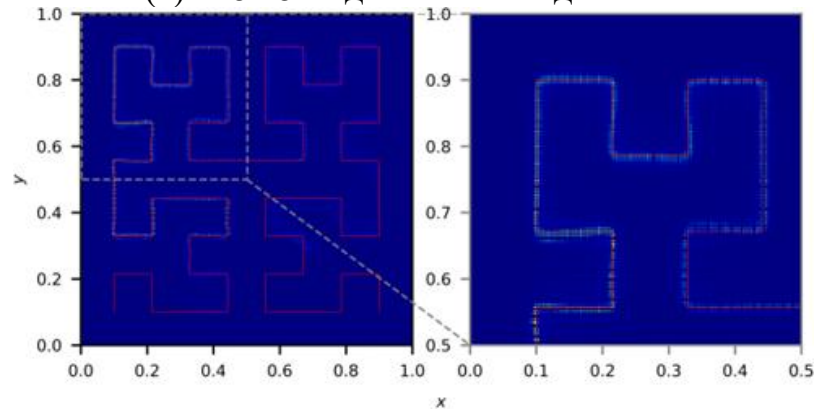
Табл. 4.1. Метрики моделей DispNet, hpDispNet, DPT та hpDPT [21]. Усі метрики наведено для моделей у форматі FP32 (виконання на CPU), моделей у форматі W8A8, що запускаються на DSP, а також моделей W8A8 у режимі запуску на CPU (режим деквантизації). Найкращі результати для DSP виділено жирним шрифтом. Результати для оригінальних моделей DispNet та DPT позначено сірим кольором.

Модель	Abs Rel, % ↓			RMSE, px ↓			S_c ↑			EPE, px ↓			D1, % ↓		
	FP32	W8A8	W8A8	FP32	W8A8	W8A8	FP32	W8A8	W8A8	FP32	W8A8	W8A8	FP32	W8A8	W8A8
	CPU	DSP	CPU	CPU	DSP	CPU	CPU	DSP	CPU	CPU	DSP	CPU	CPU	DSP	CPU
DispNet	1.01	2.03	1.15	1.12	2.24	1.10	0.86	0.58	0.68	0.29	0.69	0.31	1.81	5.35	1.73
h1DispNet	1.06	1.50	1.12	0.97	1.09	0.97	0.86	0.67	0.79	0.27	0.35	0.29	1.27	2.25	1.27
h2DispNet	0.85	0.98	0.88	0.90	0.94	0.91	0.87	0.75	0.83	0.22	0.25	0.23	1.01	1.26	1.02
h3DispNet	0.88	0.93	0.87	1.00	1.03	1.00	0.87	0.81	0.86	0.24	0.24	0.24	1.25	1.26	1.25
h4DispNet	0.90	0.94	0.92	1.02	1.02	1.02	0.85	0.83	0.85	0.24	0.25	0.24	1.24	1.25	1.24
DPT	0.75	4.18	1.48	0.87	2.09	1.03	0.89	0.52	0.87	0.21	1.03	0.39	0.95	5.78	1.39
h1DPT	0.70	1.48	0.78	0.88	1.41	0.89	0.88	0.54	0.88	0.20	0.41	0.22	1.01	2.77	1.03
h2DPT	0.71	1.12	0.72	0.91	1.02	0.91	0.88	0.62	0.88	0.20	0.29	0.20	1.07	1.27	1.08
h3DPT	0.55	1.35	0.63	0.80	1.33	0.80	0.90	0.70	0.90	0.15	0.32	0.17	0.78	1.28	0.78
h4DPT	0.74	1.27	0.76	0.94	1.38	0.94	0.87	0.73	0.86	0.21	0.32	0.21	1.14	1.62	1.14

Модифіковані моделі для всіх порядків кривих p працюють краще за оригінальні; найкращий результат досягається при $p = 3$. Як показано на Рис. 4.9, квантована модель h3DispNet зберігає здатність передбачати точки на кривій Гільберта як при запуску на CPU, так і на DSP. Модель h3DispNet на CPU демонструє майже ту ж якість, що й FP32-модель, і перевершує оригінальну DispNet за всіма метриками. На DSP погіршення якості для оригінальної DispNet є значним: AbsRel зростає з 1.01 до 2.03, а S_c знижується з 0.86 до 0.58. Модель h3DispNet майже повністю компенсує це зниження, досягаючи AbsRel 0.93 і $S_c = 0.81$.



(а) W8A8 модель на CPU делегаті



(б) W8A8 модель на DSP делегаті

Рис. 4.9. Двовимірна гістограма вихідних даних моделі h3DispNet у форматі W8A8 для CPU та DSP делегатів [21].

Для моделі DPT ситуація подібна, однак зниження якості після квантування є більш помітним як на CPU, так і на DSP. На CPU модель h3DPT демонструє кращу якість, ніж оригінальна FP32 DPT. На DSP найкращий результат показує модель h2DPT, для якої у порівнянні з оригінальною моделлю AbsRel покращується з 4.18% до 1.12%, а S_c – з 0.52 до 0.62. Крива третього порядку демонструє найкращий компроміс між покращенням AbsRel і S_c для обох моделей – DispNet і DPT. Цей висновок залишається вірним і при аналізі метрик RMSE, EPE та D1.

Якісні результати для моделі h3DPT показано на Рис. 4.10. Зменшення похибки квантування між оригінальною моделлю (Рис. 4.10, б) та модифікованою (Рис. 4.10, в) є значним, що видно на картах помилок (Рис. 4.10, д, е). Також на Рис. 4.10 (в) та Рис. 4.10 (б) видно, що h3DPT краще відтворює просторові деталі, ніж оригінальна модель. Це пояснюється підвищенням ефективної бітності для представлення карти глибини на $\log_2 L$ біт (приблизно на

2.85 біт для h3DPT, що відповідає точності INT10 – INT11). Це відображається у зростанні S_c зі збільшенням p (див. Табл. 4.1).

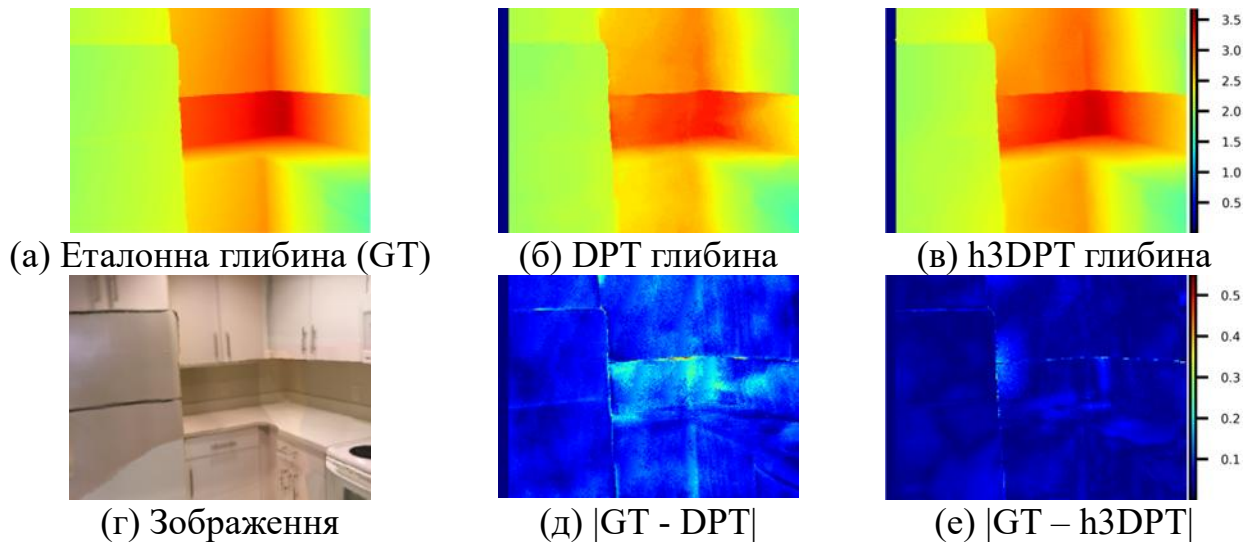


Рис. 4.10. Похибки квантування моделей DPT та h3DPT у форматі W8A8 на DSP [21].

Візуально підвищення точності добре помітно в областях з однорідними площинами, що ілюструється на Рис. 4.11.

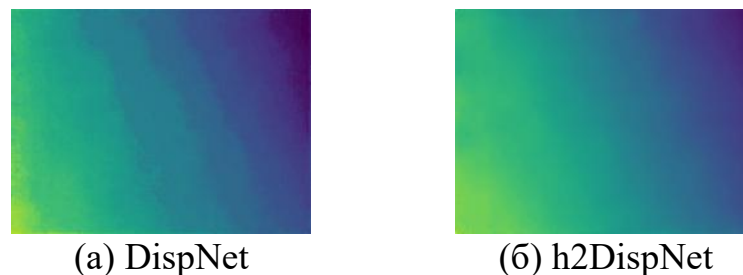


Рис. 4.11. Вплив збільшеної розрядності на якість передбачення глибини для однорідних ділянок на DSP [21].

Як видно з наведених результатів, запропонований метод не лише вирішує основну задачу підвищення розрядності, але й суттєво знижує рівень похибки квантування під час прогнозування карт глибини. Останній ефект є позитивним побічним результатом.

4.3.4. Порівняння моделей FP16, W8A16 та W8A8

Порівняно з оригінальною моделлю, накладні витрати запропонованого методу складаються з трьох частин: (1) збільшення складності моделі, (2) додаткова передача даних з DSP на CPU, (3) постобробка компонентів кривої

Гільберта. Експерименти показали, що ці накладні витрати ($\approx 14\%$) є суттєво меншими за виграш у продуктивності від використання моделей W8A8 замість W8A16 або FP16. Детальні результати порівняння для моделей з кривою третього порядку на пристрої, включаючи час виконання (T) та енергоспоживання (P) наведено в Табл. 4.2.

Табл. 4.2. Продуктивність оригінальних і модифікованих моделей [21].

Точність	Abs Rel, %	EPE, пкс	D1, %	S_c	T, мс	P, мВт · с/infr.
DispNet						
FP32	1.01	0.29	1.81	0.858	-	-
FP16	1.50	0.37	1.80	0.855	19.54	19.52
W8A16	1.78	0.63	5.22	0.798	18.7	12.3
W8A8	2.02	0.69	5.34	0.585	10.5	7.1
Запропонована W8A8	0.93	0.24	1.26	0.807	12.0	8.7
DPT						
FP32	0.75	0.21	0.95	0.889	-	-
FP16	1.14	0.27	0.97	0.884	54.1	110.5
W8A16	4.03	0.97	5.58	0.825	46.2	64.5
W8A8	4.16	1.03	5.76	0.520	26.7	28.3
Запропонована W8A8	1.35	0.32	1.28	0.697	30.4	29.7

Вимірювання для запропонованого методу враховують передачу даних з DSP на CPU та постобробку компонент кривої Гільберта на CPU. Накладні витрати запропонованого методу проявляються у збільшенні часу виконання та енергоспоживання між модифікованими та оригінальними моделями у форматі W8A8. Моделі у форматі FP32 запускаються на CPU; моделі у форматах FP16, W8A16, W8A8 – на DSP.

Модифікована модель DispNet у форматі W8A8 демонструє кращу якість, ніж оригінальна модель у форматі W8A16, одночасно зменшуючи енергоспоживання на 35% та затримку на 30%. Порівняно з оригінальною моделлю у форматі W8A16, модифікована модель DPT у форматі W8A8 показує суттєво кращий показник AbsRel і лише трохи гірший S_c . Водночас знижуються енергоспоживання на 34% і час обробки на 54%.

На Рис. 4.12 – Рис. 4.16 наведено додаткові приклади карт глибини, передбачених оригінальними та модифікованими моделями DispNet і DPT. Ці приклади охоплюють як простіші сцени (Рис. 4.12 – Рис. 4.15), так і складнішу

сцену (Рис. 4.16). У всіх прикладах модифіковані моделі демонструють значно меншу похибку квантування; залишкові похибки зосереджуються переважно в областях розривів глибини. Похибки поблизу розривів глибини також спостерігаються для FP32-моделей і не пов'язані з запропонованим підходом.

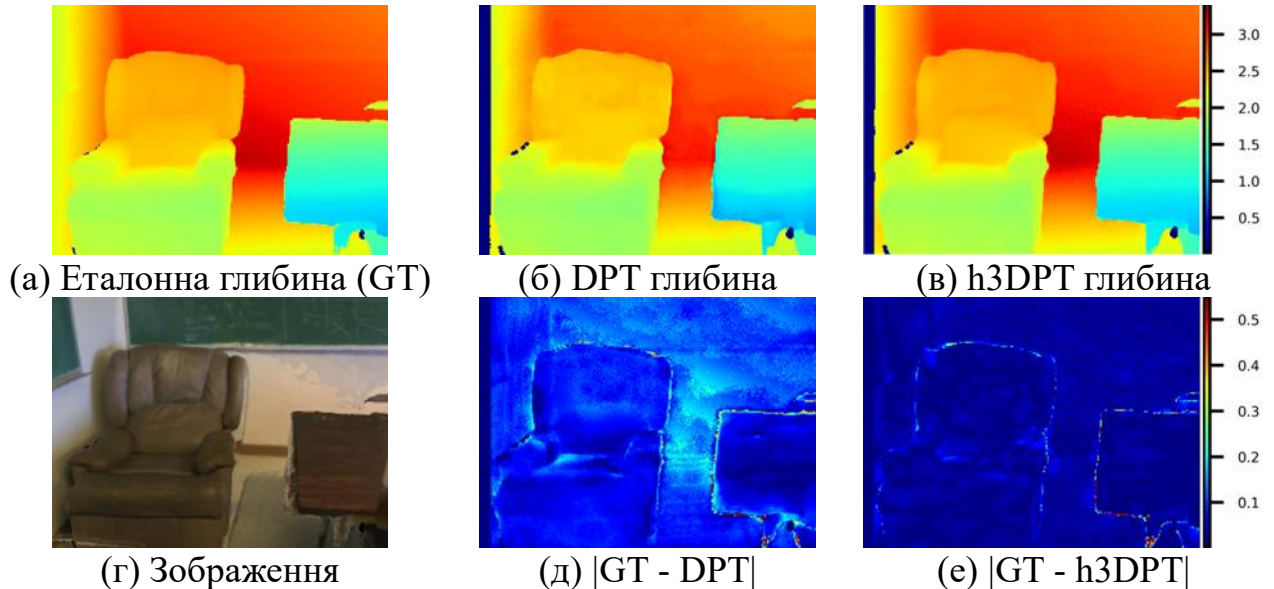


Рис. 4.12. Похибки глибини для моделей DPT та h3DPT на DSP [21]. Сцена scene0030_02 набору даних ScanNet. Усі значення подано в метрах.

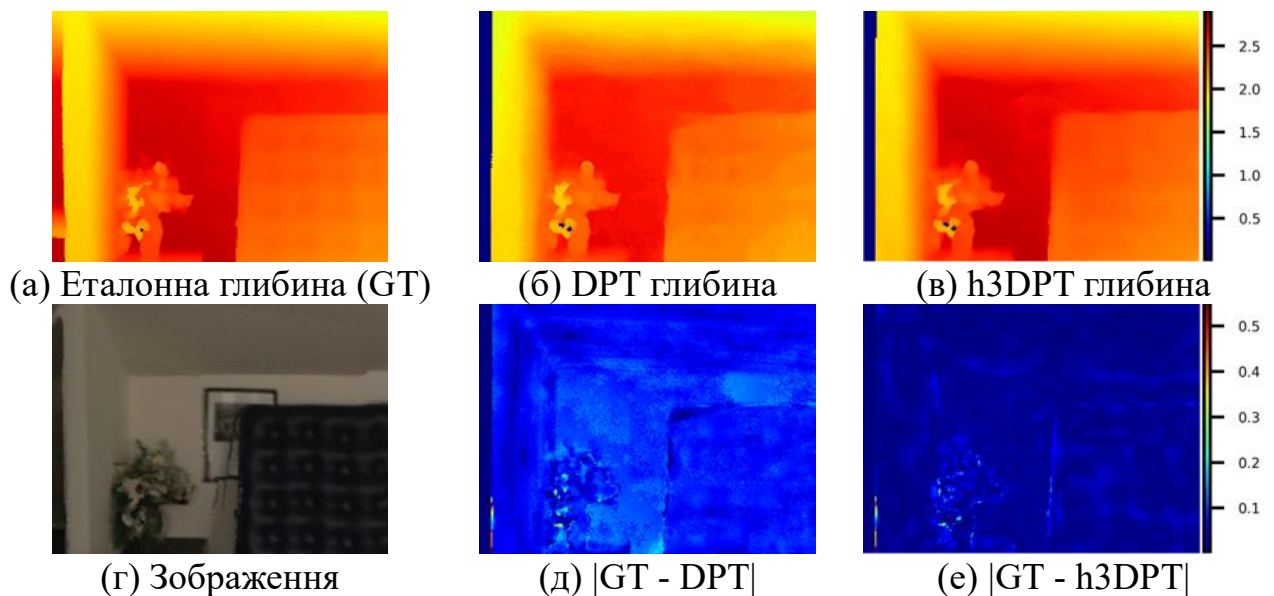


Рис. 4.13. Похибки глибини для моделей DPT та h3DPT на DSP [21]. Сцена scene0629_00 набору даних ScanNet. Усі значення подано в метрах.

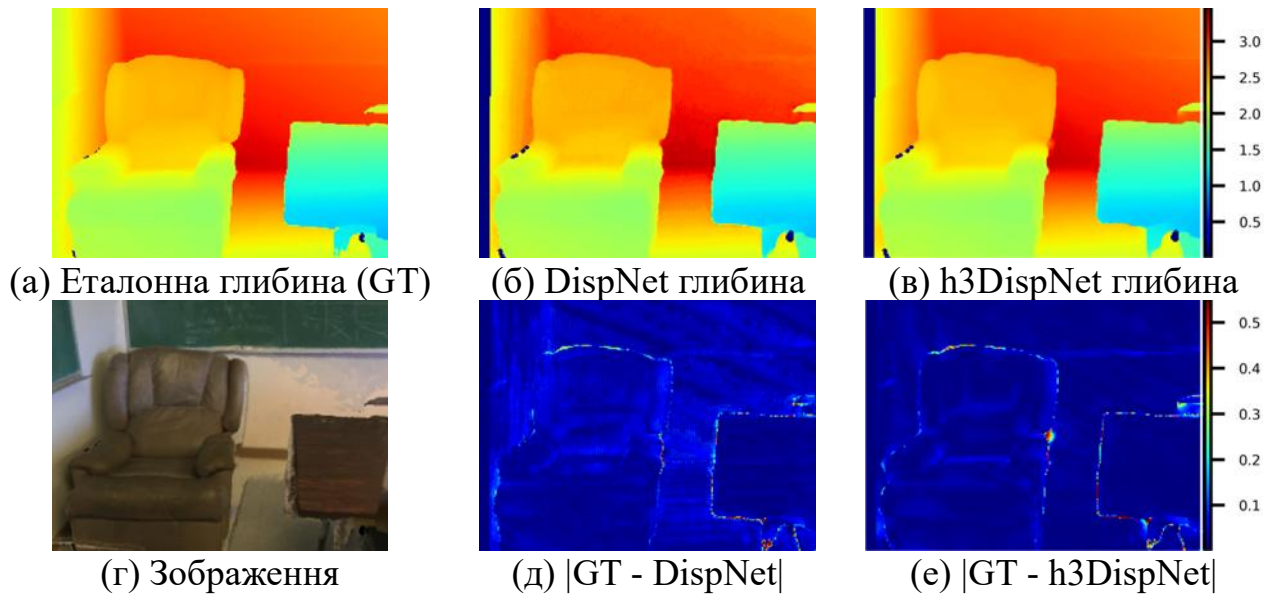


Рис. 4.14. Похибки глибини для моделей DispNet та h3DispNet на DSP [21].
Сцена scene0030_02 набору даних ScanNet. Усі значення подано в метрах.

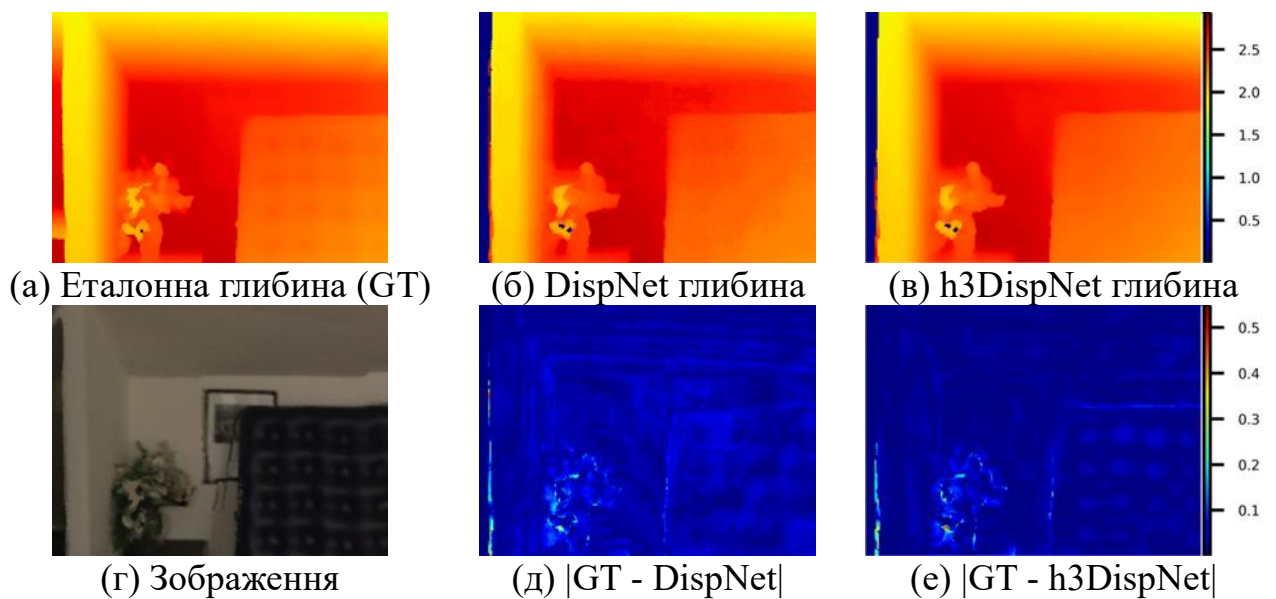


Рис. 4.15. Похибки глибини для моделей DispNet та h3DispNet на DSP [21].
Сцена scene0629_00 набору даних ScanNet. Усі значення подано в метрах.

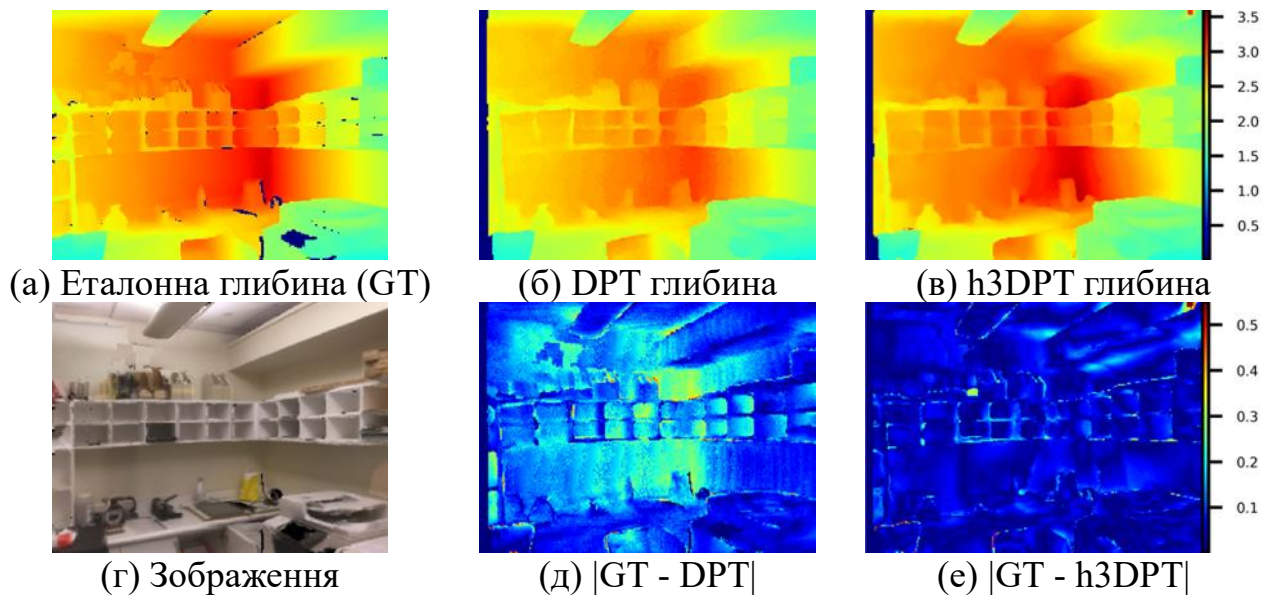


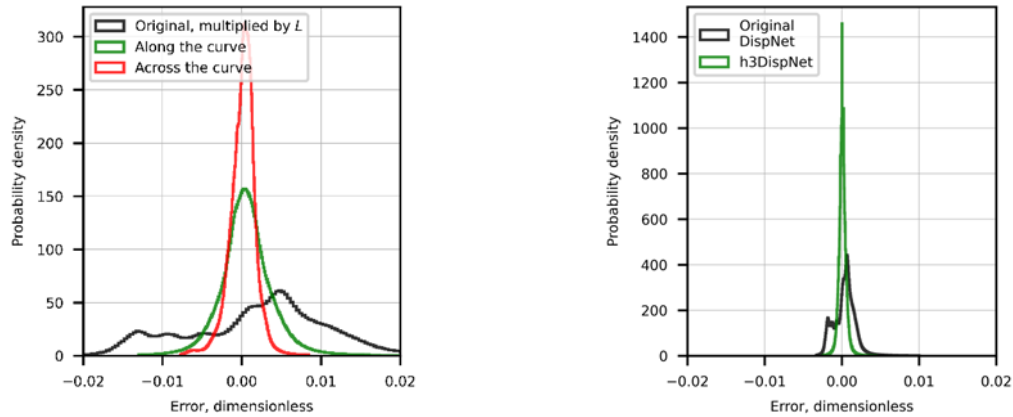
Рис. 4.16. Похибки глибини для моделей DPT та h3DPT на DSP [21]. Сцена scene0804_00 набору даних ScanNet. Усі значення подано в метрах.

4.3.5. Аналіз зменшення помилок квантування

Для детального аналізу зменшення помилок квантування було обрано h3DispNet модель. Щоб виключити вплив втрат точності обчислень, аналіз проведено для моделі W8A8, яка запускалася на CPU. Основна відмінність між оригінальною та модифікованою моделлю полягає у кількості використаних каналів: взаємозв'язок між похибками квантування компонентів кривої Гільберта потребує подальшого вивчення. Початковий аналіз базувався на припущенні незалежності помилок квантування між компонентами. Проте оцінка на реальних даних показала, що ця гіпотеза не завжди вірна, оскільки було виявлено кореляцію вздовж кривої Гільберта. Кількісно, розподіл помилок уздовж кривої ширший, ніж поперек неї (див. Рис. 4.17, а).

Похибки поперек кривої здебільшого нівелюються під час постобробки, тоді як уздовж-кривої – стискаються в L разів і визначають рівень похибки цільової величини – диспаратності. Для порівняння помилок у єдиній шкалі, на Рис. 4.17 (а) показано похибку диспаратності оригінальної моделі, помножену на довжину кривої L . Похибки квантування у представленні на основі кривої Гільберта суттєво менші, ніж очікувані при масштабуванні похибки оригінальної моделі в L разів. Таким чином, у просторі диспаратності похибка квантування

модифікованої моделі значно менша, ніж у оригінальній моделі (див. Рис. 4.17, б): $\hat{\sigma}$ дорівнює $5.65 \cdot 10^{-4}$ та $17.66 \cdot 10^{-4}$ відповідно. Отже, ми отримали зменшення похибки квантування приблизно у 3.1 рази на CPU. На DSP це покращення становить близько 4.6 рази. Подібний ефект спостерігається і для DPT моделі.



(а) Розподіли похибок вздовж та поперек кривої Гільберта.

(б) Розподіл похибок диспаратності для моделей h3DispNet та DispNet.

Рис. 4.17. Розподіли похибок квантування для моделі h3DispNet [21]. Значення диспаратності d обчислюються з компонент кривої Гільберта x та y і нормалізуються до діапазону $[0,1]$.

4.3.6. Експеримент на наборі даних KITTI 2012

З метою аналізу можливості застосування запропонованого підходу для задачі реконструкції карт глибини з іншого домену було проведено експеримент на наборі даних KITTI 2012 [164]. KITTI 2012 – це набір реальних даних з домену автономного водіння з розрідженими еталонними значеннями диспаратності, що були отримані за допомогою системи LiDAR. Для оцінювання використано 194 зображення з тренувальної частини набору KITTI 2012. Навчальний набір даних сформовано з поєднання ScanNet та Virtual KITTI 2 [170] із балансуванням 25/75%. Моделі DispNet і h2DispNet навчалися з роздільною здатністю вхідних зображень 256×1152 пікселів та роздільною здатністю вихідних карт диспаратності 128×576 пікселів. Параметри навчання такі ж, як в експерименті на ScanNet.

Табл. 4.3 Результати на наборі даних KITTI 2012 для моделей DispNet та h2DispNet [21]. Моделі у форматі FP32 виконуються на CPU; моделі у форматі W8A8 – на DSP. Найкращі метрики для моделей у форматі W8A8 виділено жирним шрифтом.

Точність	Abs Rel, % ↓	EPE, пікс ↓	D1, % ↓
FP32	3.88	1.38	9.13
W8A8	5.01	1.63	9.95
Запропонована, FP32	3.24	1.07	5.22
Запропонована, W8A8	3.30	1.09	5.36

Як показано в Табл. 4.3, для оригінальної моделі DispNet досягнуто значень EPE 1.38 пікс. і D1 9.13%. Цікаво, що модель h2DispNet демонструє кращі результати – EPE 1.07 пікс. і D1 5.22%. Метрика S_c виключена з аналізу, оскільки її не можна коректно застосовувати до розріджених еталонних значень диспаратностей набору KITTI 2012. Модель DispNet, квантована до формату W8A8, демонструє погіршення якості при запуску на DSP: EPE зростає до 1.63 пікс., а D1 – до 9.95%. Водночас модель h2DispNet у форматі W8A8 на DSP показує лише незначне зниження якості. Виміряна як EPE між диспаратностями моделей FP32 і W8A8, похибка квантування становить 1.01 пікс. для DispNet та 0.38 пікс. для h2DispNet. Це відповідає підвищенню якості передбачення диспаратності на пристрої у 2.6 рази. Загалом на DSP модель h2DispNet покращує показник D1 приблизно на 4.6% та EPE приблизно на 33% порівняно з оригінальною DispNet. Приклади передбачених карт диспаратності наведені на Рис. 4.18. Цей експеримент демонструє, що запропонований підхід може бути успішно застосований до реальних наборів даних з різних доменів.

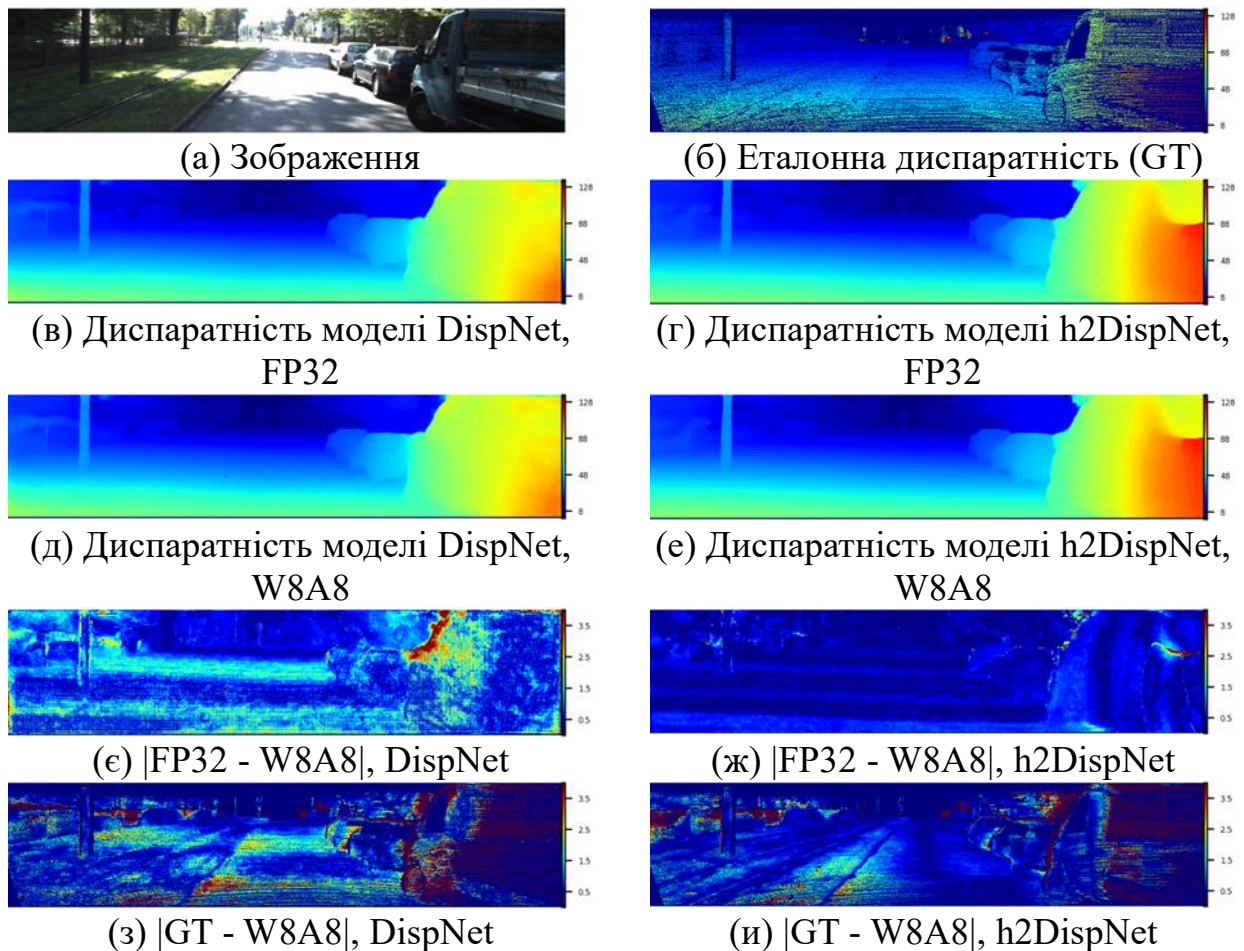


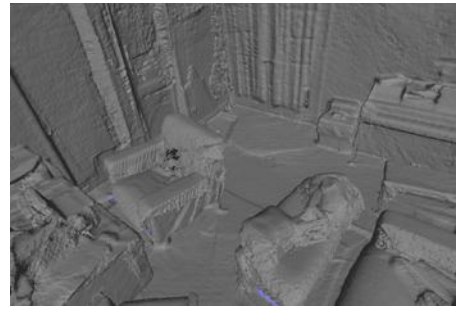
Рис. 4.18. Порівняння моделей DispNet та h2DispNet на наборі даних KITTI 2012 [21].

Різниця між картами диспаратності, передбаченими моделями у форматі FP32 (CPU) (Рис. 4.18, в та Рис. 4.18, г) та W8A8 (DSP) (Рис. 4.18, д та Рис. 4.18, е), показано на зображеннях Рис. 4.18 (є) та Рис. 4.18 (ж); абсолютні похибки передбачення диспаратності для моделей у форматі W8A8 (DSP) наведено на зображеннях Рис. 4.18 (з) та Рис. 4.18 (и).

4.3.7. Вплив якості квантування на 3D реконструкцію

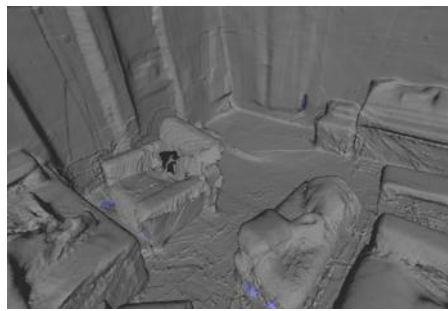
Проведено додатковий експеримент, щоб з'ясувати, як артефакти квантування в картах глибини впливають на 3D реконструкцію сцени. Для об'єднання карт глибини в 3D меш використано підхід на основі підходу TSDF [171], що реалізований у Open3D бібліотеці Python [172]. Застосовано масштабований TSDF об'єм із параметрами $voxel_length = 0.01m$, $sdf_trunc = 0.15$. Для експериментів обрано сцену scene0050_02 з набору ScanNet, що складається

з 4379 кадрів. Для 3D реконструкції використовувався кожен 40-й кадр. Якість мешу, отриманого з еталонних карт глибини, показано на Рис. 4.19.

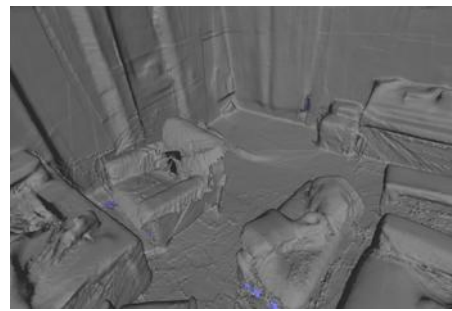


(а) Еталонний (GT) меш з текстурою (б) Еталонний (GT) меш без текстури
Рис. 4.19. Вигляд 3D-моделі, отриманої за рахунок об'єднання еталонних (GT) карт глибини для сцени scene0050_02 набору даних ScanNet [21].

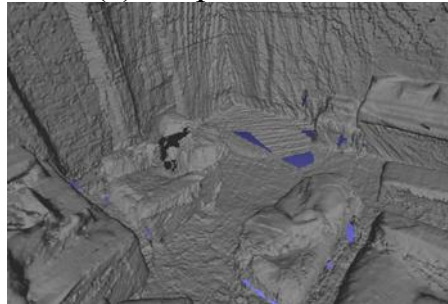
На Рис. 4.20 та Рис. 4.21 наведено якісні результати для моделей h2DispNet, h2DPT та їх базовими варіантами відповідно. 3D реконструкції, побудовані за картами глибини, передбаченими моделями FP32 h2DispNet (Рис. 4.20, б) та FP32 h2DPT (Рис. 4.21, б), мають подібну структуру та плавність до моделей FP32 DispNet (Рис. 4.20, а) і FP32 DPT (Рис. 4.21, а). І базові, і модифіковані FP32 моделі забезпечують якість 3D реконструкції, близьку до якості отриманої при об'єднанні еталонних карт глибини (Рис. 4.19), хоча й з дещо більшою згладженістю.



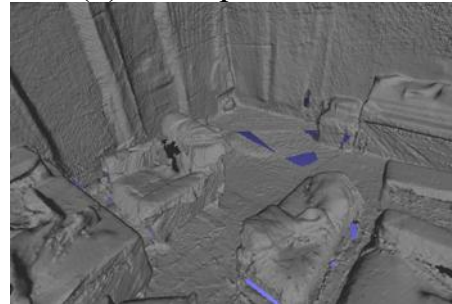
(a) DispNet, FP32



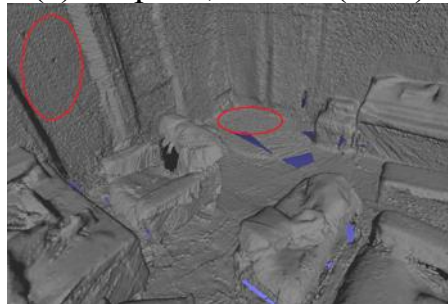
(б) h2DispNet, FP32



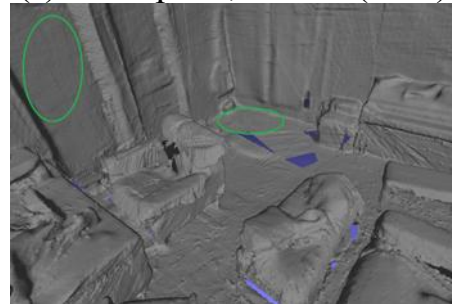
(в) DispNet, W8A8 (DSP)



(г) h2DispNet, W8A8 (DSP)



(д) DispNet, W8A8 (CPU)



(е) h2DispNet, W8A8 (CPU)

Рис. 4.20. Вигляд 3D моделі, зібраної на основі передбачених карт глибини від моделей DispNet та h2DispNet для сцени scene0050_02 з набору даних ScanNet [21]. Деякі похибки реконструкції позначено червоним, а покращені структури – зеленим.

Помічено два типи артефактів квантування, присутніх у 3D реконструкції, побудованої з карт глибини моделей, що запускались на CPU. Перший тип – шум на плоских поверхнях для оригінальної моделі DispNet у форматі W8A8 (Рис. 4.20, д); другий – наявність помітних меж між глибинами різних кадрів (схожі на сходи) у реконструкції для оригінальної моделі DPT у форматі W8A8 (Рис. 4.21, д). Останній тип артефактів викликаний систематичними похибками у передбаченні глибини, які призводять до помилок у масштабі глибини. Моделі, модифіковані відповідно до запропонованого методу, призводять до значно меншого рівня шуму у відновленій 3D моделі (Рис. 4.20 (е) та Рис. 4.21 (е)).

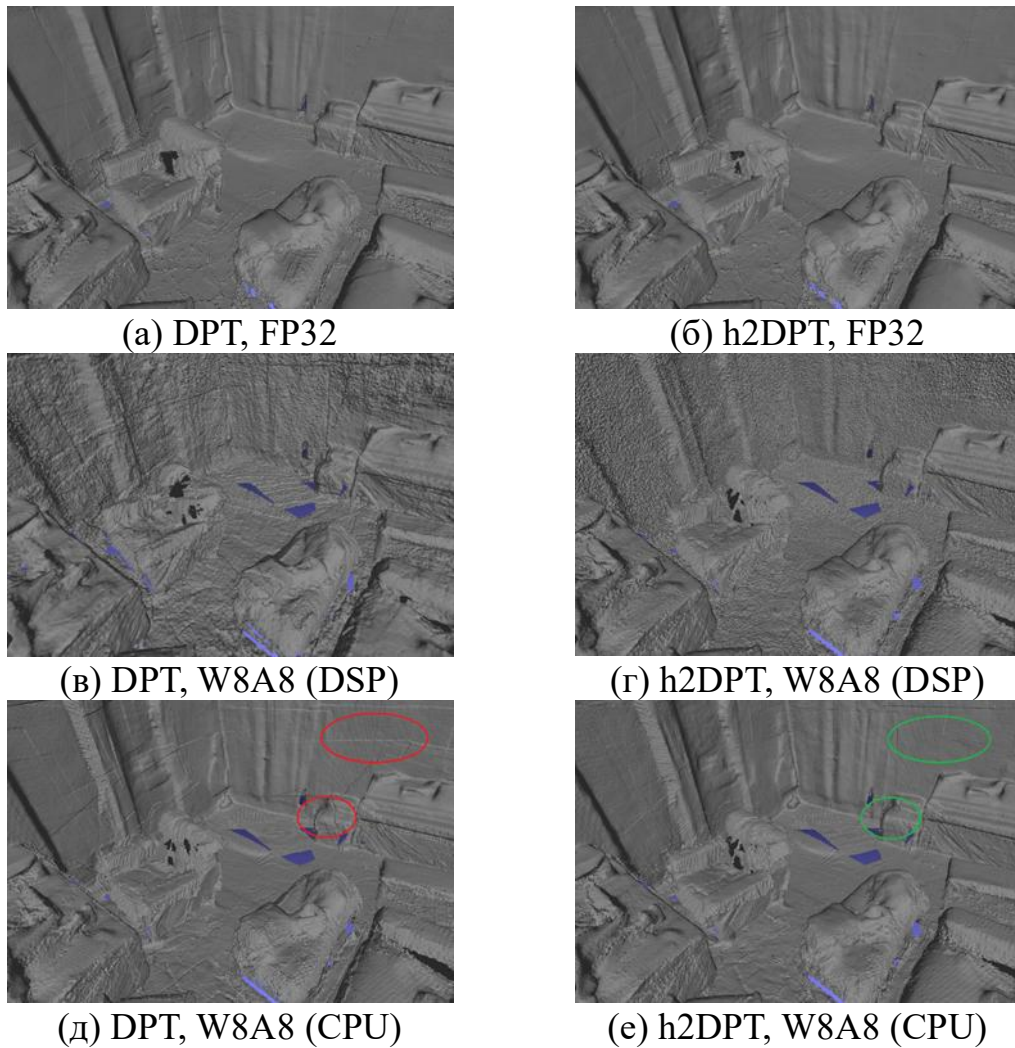


Рис. 4.21. Вигляд 3D моделі, зібраної на основі передбачених карт глибини від моделей DPT та h2DPT для сцени scene0050_02 набору даних ScanNet [21]. Деякі похибки реконструкції позначено червоним, а покращені структури – зеленим.

Для базових моделей виконаних на DSP, спостерігаються ті ж артефакти, але більш виражені (Рис. 4.20, в та Рис. 4.21, в). Модель h2DispNet у форматі W8A8 практично усуває артефакти квантування (Рис. 4.20, г) і відновлює просторові деталі 3D реконструкції. Модель h2DPT у форматі W8A8 прибирає артефакти типу «сходинок» та зменшує шум на плоских поверхнях.

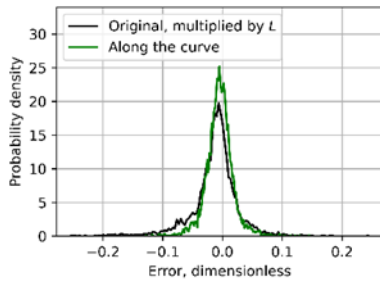
4.3.8. Експеримент з оцінювання пози людини

З метою аналізу можливості застосування запропонованого підходу до іншою задачі комп'ютерного зору було проведено експеримент з оцінювання пози людини (HPE). Були реалізовані два підходи: пряме регресування ключових

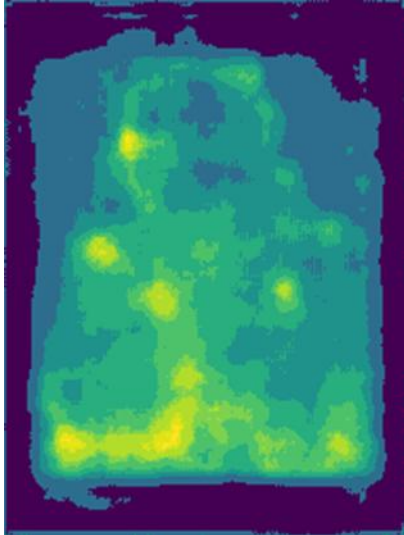
точок [173] та представлення ключових точок у вигляді теплових карт [174]. Для навчання використовувався набір даних MS COCO [175] з анотаціями ключових точок пози людини. Обидві моделі мають енкодер ResNet-RS [176] з параметром α , що дорівнює 0.5. Голова для прямого регресування складається з повнозв'язних шарів, як запропоновано в роботі [173]. Модель для передбачення теплових карт використовує декодер DispNet та згорткову голову, описану в роботі [174].

У цьому експерименті були навчені дві легкі моделі: для прямого передбачення ключових точок з $AP_{50} = 43$ та на основі теплових карт з $AP_{50} = 83$. Виявлено, що найстабільніший спосіб навчання модифікованої НРЕ-моделі полягає у попередньому навчанні оригінальної моделі, подальшому використанні її як вчителя для обчислення компонентів кривої Гільберта x_{GT} та y_{GT} з її прогнозів у функції втрат (4.3). Такий підхід дозволяє навчати модифіковану модель навіть за відсутності еталонних даних для обчислення x_{GT} та y_{GT} , наприклад, у випадку навчання теплових картах.

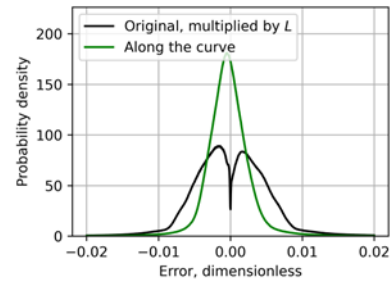
Для моделі з прямим передбаченням ключових точок для моделі на основі кривої Гільберта похибка квантування вздовж кривої при $p=1,2$ зростає приблизно у L разів порівняно з похибкою квантування ключових точок оригінальної моделі (див. Рис. 4.22 (а) для випадку $p=1$); похибка поперек кривої є на порядок меншою та відповідає похибці квантування оригінальної моделі. У результаті загальна похибка квантування в просторі ключових точок для модифікованої моделі залишається майже незмінною. Це свідчить про те, що для задачі прямої регресії похибка квантування є сильно залежною від сигналу. Єдиний ефект від запропонованої модифікації полягає в підвищенні точності вихідного представлення (кількості бітів).



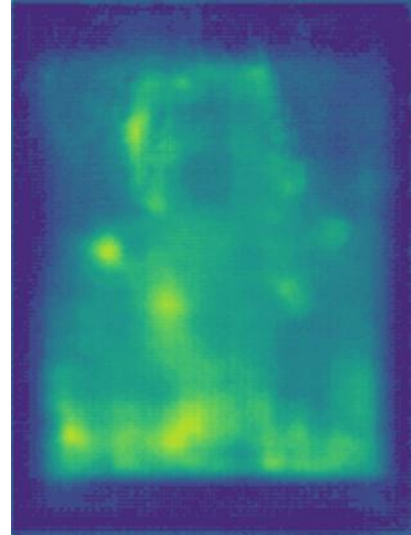
(а) Похибка ключових точок



(в) Теплова карта, отримана за допомогою оригінальної моделі НРЕ



(б) Похибка теплових карт



(г) Теплова карта, отримана за допомогою модифікованої моделі НРЕ

Рис. 4.22. Похибка квантування для НРЕ моделі при прямій регресії ключових точок (а) та теплових картах (б). Порівняння передбачення теплових карт на DSP за допомогою моделей W8A8 (в, г). [21]

Для моделі передбачення теплових карт ситуація відрізняється. У цьому випадку похибка квантування вздовж кривої Гільберта другого порядку зростає менш суттєво порівняно з оригінальною моделлю. Як видно з Рис. 4.22 (б), розподіл похибки вздовж кривої є вужчим, ніж похибка квантування оригінальних теплових карт, помножена на довжину кривої. У результаті похибка квантування в просторі теплових карт для модифікованої моделі в 2.69 рази менша, ніж для оригінальної моделі. Крім того, покращується просторове відображення передбачених теплових карт, що показано на Рис. 4.22 (в) та Рис. 4.22 (г). Цей ефект також пов'язаний із підвищенням точності (розрядності) вихідного представлення.

4.4. Обмеження підходу

Запропонований метод розроблений з основною метою – підвищити розрядність передбаченої карти глибини на пристроях з обмеженою точністю обчислень. Крім цього, спостерігався позитивний побічний ефект, у вигляді суттєвого зменшення похибки квантування. Цей ефект проявляється в різній мірі для задач прогнозування глибини та оцінки пози людини за тепловими картами (додатковий експеримент), але не спостерігається для оцінки пози людини за допомогою прямої регресії ключових точок (додатковий експеримент).

Експерименти показують, що похибки квантування компонентів кривої Гільберта можуть бути корельованими та мати значне стандартне відхилення вздовж кривої. Цей ефект яскраво проявляється для задачі оцінки пози людини з прямою регресією та лише незначною мірою для моделей стереозіставлення. Така кореляція свідчить про те, що модифікована модель замість навчання незалежних гілок для передбачення компонентів кривої Гільберта, формує внутрішнє представлення цільової величини (наприклад, диспаратності, глибини або координат ключових точок), яке на останніх шарах перетворюється у компоненти кривої Гільберта. У такому випадку модель передбачає ту саму величину з аналогічною похибкою квантування, що й оригінальна модель, але ця похибка проходить пряме перетворення до Гільбертового представлення (всередині моделі) і зворотне перетворення (на етапі постобробки). У результаті рівень похибки залишається незмінним, і ефект від застосування запропонованого методу полягає лише в підвищенні розрядності. Це характерно для НРЕ з прямою регресією.

Якщо компоненти x та y передбачаються незалежними гілками моделі, то їхні похибки квантування компонент кривої Гільберта не є корельованими та мають схожий рівень, як і в оригінальній моделі. Ця похибка стискається в L разів на етапі постобробки. У такому випадку запропонований метод дозволяє як зменшити похибку квантування, так і збільшити розрядність, як було продемонстровано в експериментах зі стереозіставлення. Якщо вдасться виявити

та усунути джерело кореляції вздовж кривої, метод може бути застосований для зменшення похибки квантування навіть у випадках, коли підвищення розрядності не є критичним.

Незважаючи на покращення якості передбачення глибини, запропонований метод вимагає повторного навчання моделі з високою точністю. Крім того, необхідно, щоб похибки квантування були обмеженими, оскільки великі значення похибки порушують взаємно однозначне відображення між значенням глибини та точками на кривій Гільберта. Наступним кроком є об'єднання запропонованого підходу з квантуванням під час навчання (QAT) і перехід до вищих розмірностей, де багатовимірною кривою Гільберта тієї ж довжини заповнює простір менш щільно.

У даній роботі наведено результати для моделей стереозіставлення, які передбачають диспаратність в один етап. Новітні моделі, що демонструють найвищу якість у стереозіставленні [136, 177] та монокулярному прогнозуванні глибини [178], використовують ітеративне уточнення диспаратності за допомогою GRU або інших рекурентних блоків. У таких архітектурах диспаратність покращується протягом кількох ітерацій шляхом додавання невеликого корегуючого сигналу до карти глибини прогнозованої попередньо. Інтеграція запропонованого підходу до представлення диспаратності як компонентів кривої Гільберта в ітеративні моделі суттєво відрізняється від його використання в одноетапних моделях, таких як DispNet або DPT. Хоча і було проведено ряд експериментів з модифікацією представлення вихідних даних, повна інтеграція до ітеративних архітектур вимагає глибшого втручання в модель на кількох рівнях. Дана ідея залишається як перспективний напрямок для подальших досліджень.

4.5. Висновки до розділу

У цьому розділі запропоновано новий підхід до передбачення глибини з високим динамічним діапазоном на пристроях із низькою розрядністю, який ґрунтується на представленні глибини у вигляді точок на 2D кривій Гільберта.

Таке представлення фактично кодує глибину з великим динамічним діапазоном як дві компоненти кривої Гільберта з низьким діапазоном. Модель передбачення карт глибини навчається безпосередньо прогнозувати компоненти кривої Гільберта, які обчислюються на пристрої з низькою розрядністю і використовуються для відновлення глибини з високою розрядністю. Крім підвищення розрядності, запропонований метод дозволяє зменшити похибку квантування до 4.6 разів. Експерименти показали, що для задачі стереозіставлення запропонований метод дозволяє реконструювати глибину з точністю INT10–INT11 для моделі, квантованої у форматі W8A8, з якістю, подібною або кращою, ніж у оригінальній моделі у форматі W8A16. Таким чином, прогноз карт глибини на пристрої виконується без втрати якості у 1.4–2 рази швидше та з 65% зниженням енергоспоживання. Запропонований підхід є корисним для застосування на пристроях у різних задачах передбачення карти щільної глибини, включно з монокулярним і стерео-прогнозуванням, Multi-View Stereo, доповненням глибини, покращенням її якості та інпейнтингом. Подальші дослідження мають бути спрямовані на глибше розуміння і повне вивчення ефекту зменшення похибки квантування.

Основні внески описані в цьому розділі можна узагальнити наступним чином:

1. Запропоновано новий підхід до якісного прогнозування глибини на пристроях з низькорозрядною арифметикою. Метод передбачає представлення карти глибини як компонент кривої Гільберта, навчання точної моделі для їх прогнозування та реконструкцію карти глибини високої точності за допомогою компонент кривої Гільберта низької розрядності.
2. Оцінено ефективність методу і показано, що модифікована модель, квантована до формату W8A8 (8-бітні ваги та активації), може досягти подібної або кращої якості прогнозування, ніж оригінальна модель у форматі W8A16. При цьому час виконання та енергоспоживання

W8A8-моделі з постобробкою в 1.5 рази менші, ніж у вихідної W8A16-моделі.

3. Використання запропонованого методу також позитивно впливає на якість квантування зменшуючи її похибку до 4.6 разів.

ВИСНОВКИ

Дисертацію присвячено підвищенню якості, стійкості та енергоефективності методів 3D-реконструкції середовища, що виконуються безпосередньо на кінцевому пристрої користувача, з метою забезпечення реалістичної та надійної взаємодії з віртуальними об'єктами в системах доповненої реальності. Мотивацією роботи є швидке зростання застосувань сценаріїв ДР на мобільних і носимих платформах, де поєднання вимог до геометричної точності, фотореалістичності, роботи в реальному часі та суворих обмежень обчислювальних ресурсів/енергоспоживання залишається невирішеним комплексним викликом.

1. Проведено аналіз класичних та сучасних методів реконструкції карт глибини та 3D реконструкції середовища, визначено їхні обмеження для задач доповненої реальності на кінцевих пристроях користувача. Систематизовано класифікацію класичних і сучасних підходів (SfM/MVS, стерео, нейромережеві методи, NeRF-подібні представлення) саме в контексті кінцевих пристроїв ДР. Узгоджено набір критеріїв порівняння для сценаріїв ДР на кінцевому пристрої користувача, а саме: точність/стійкість \leftrightarrow час виконання/енергоефективність \leftrightarrow обмеження пам'яті/розрядності), що формує методичну основу для подальших рішень у роботі. Показано, що наявні методи добре працюють у статичних сценах з дифузними поверхнями, однак деградує за динамічного освітлення та на відбивних/напівпрозорих матеріалах. Більшість рішень орієнтовано на серверні/GPU умови, не на DSP/NPU кінцевих пристроїв. Обґрунтовано потребу адаптації наборів даних під конкретні конфігурації камер та домени застосування, а також необхідність енергоефективних/квантованих обчислень без втрати точності.

2. Удосконалено метод Neural Radiance Fields (NeRF) до умов динамічного освітлення шляхом модифікації функцій втрат та введення часової змінної, що дозволило покращити якість реконструкції сцен з динамічним освітленням та розширювати існуючі набори даних для складних сцен. Доведено, що стандартний NeRF (із просторовою позицією та напрямом погляду) не

відтворює коливання освітленості. Запропоноване введення часу та уточнена функція втрат підвищують фотореалістичність (особливо для тонких структур) і зменшують відносну похибку глибини на **10–28%**. Головний акцент зроблено на практичному використанні: підхід слугує інструментом **доповнення й адаптації наборів даних** під цільові камери/домени кінцевих пристроїв для складних умов освітлення, унаслідок чого зростає якість навчання моделей глибини, критична для сценаріїв ДР на кінцевих пристроях користувача. Водночас збережено відомі обмеження сцено-специфічного навчання NeRF, що окреслює напрями майбутньої оптимізації.

3. Вперше розроблено та запатентовано метод реконструкції глибини, який враховує напівпрозорі та відбивні поверхні, зберігає окремо значення глибини до самої площини та до відбитого/перекритого об'єкта, що дозволяє збільшити точність реконструкції складних сцен, які містять недифузні поверхні. Запропонований підхід усуває типові похибки реконструкції карт глибини та 3D реконструкції на дзеркалах, полірованих матеріалах та напівпрозорих поверхнях: метод продукує багат шарову чи багатоканальну глибину, що робить 3D-модель повнішою та фізично узгодженою. Таким чином, метод **підвищує надійність сприйняття сцен** реального світу, критичну для імерсивних сценаріїв ДР. Показано практичний вплив у низці сценаріїв, зокрема для області ДР. Патент методу реконструкції глибини, що враховує напівпрозорі та відбивні поверхні для покращення реконструкції складних об'єктів захищає запропонований метод від копіювання, збільшує патентне портфоліо компанії, де були впроваджені результати, у відповідному домені, а також, потенційно може генерувати додатковий дохід для компанії у разі виявлення використання методу конкурентами.

4. Вперше сформульовано та розроблено метод для прогнозування карт глибини з представленням виходу моделі у вигляді компонент двовимірної кривої Гільберта, що дозволило для квантованих моделей розширити діапазон глибини, підвищити її точність та покращити енергоефективність реконструкції сцени на апаратних прискорювачах з

обмеженою розрядністю (DSP/NPU) кінцевих пристроїв користувача. Запропоновано кодувати високодинамічну глибину двома малодіапазонними компонентами кривої Гільберта та **безпосередньо навчати** модель передбачати ці компоненти у форматі W8A8. Це **зменшує похибку квантування до 4.6 разів**, забезпечує ефективну реконструкцію глибини з еквівалентною точністю INT10–INT11 для квантованих моделей і на практиці дає **1.4–2× прискорення при ≈65% зниженні енергоспоживання** порівняно з W8A16, зберігаючи або покращуючи якість. Метод є універсальним для монокулярної/стерео-глибини, MVS, доповнення, підвищення якості та інпейнтингу; окреслено подальші дослідження щодо повного пояснення ефекту зменшення похибки квантування.

5. Результати присутні в даному дослідженні були використані в рамках комерційних та науково-дослідницьких проектів ТОВ «Самсунг РнД Інститут Україна». Розглянуті та запропоновані методи 3D реконструкції сцени на кінцевих пристроях користувача з обмеженими обчислювальними ресурсами знайшли застосування у галузі візуального інтелекту та були використані при розробці комерційних проектів, що спрямовані на сценарії створення/редагування просторового контенту для флагманської моделі смартфона Samsung Galaxy S25.

У роботі виконано послідовний перехід від системного аналізу вимог сценаріїв ДР на кінцевих пристроях споживача щодо задач реконструкції карт глибини та 3D реконструкції середовища, до трьох взаємодоповнювальних технічних рішень: 1) Розширення наборів даних, зокрема для складних умов динамічного освітлення за рахунок використання адаптованого NeRF із часовою змінною та функцією втрат за глибиною; 2) Геометрично та фізично обґрунтованого відновлення складних поверхонь (багатошарова глибина для відбивних/напівпрозорих об'єктів); 3) алгоритмічно-апаратної оптимізації передбачення глибини (подання через криву Гільберта для квантованих DSP/NPU). Сукупно це забезпечує підвищення точності та стійкості 3D-реконструкції за реальних складних умов та істотне поліпшення продуктивності й енергоефективності на кінцевих пристроях. Практичне значення результатів

полягає у можливості інтеграції запропонованих рішень у новітні методи та підходи орієнтовані на сценарії доповненої реальності на кінцевий пристрій користувача, а також в галузі робототехніки, автономної навігації, та ін.

Таким чином, запропоновані нові методи та підходи реконструкції карт глибини та 3D реконструкції сцени дозволяють підвищити якість, стійкість та енергоефективності методів 3D реконструкції середовища, що виконуються на кінцевому пристрої користувача, для забезпечення реалістичної та надійної взаємодії користувача з віртуальними об'єктами в системах доповненої реальності.

Запропоновані підходи можна приміняти не тільки для ДР, а також до таких галузей як автономне водіння, робототехніка, інтернет речей, безпілотники, та інші області, які потребують ефективного визначення карт глибини на кінцевому пристрої, що робить дослідження вагомим внеском у розвиток методів прикладної математики та обробки зображень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Stirenko, S., Kochura, Y., Alienin, O., Rokovyi, O., Gang, P., Zeng, W., Gordienko, Y.: Chest X-Ray Analysis of Tuberculosis by Deep Learning with Segmentation and Augmentation. (2018).
<https://doi.org/10.48550/ARXIV.1803.01199>
2. Gordienko, Yu., Gang, P., Hui, J., Zeng, W., Kochura, Yu., Alienin, O., Rokovyi, O., Stirenko, S.: Deep Learning with Lung Segmentation and Bone Shadow Exclusion Techniques for Chest X-Ray Analysis of Lung Cancer. (2017).
<https://doi.org/10.48550/ARXIV.1712.07632>
3. Alekseichuk, L., Lande, D., Novikov, O.: Application of Large Language Models for Assessing Parameters and Possible Scenarios of Cyberattacks on Information and Communication Systems. Theor. Appl. Cybersecurity. 6, (2024).
<https://doi.org/10.20535/tacs.2664-29132024.1.315242>
4. Stirenko, S., Gordienko, Yu., Shemsedinov, T., Alienin, O., Kochura, Yu., Gordienko, N., Rojbi, A., Benito, J.R.L., González, E.A.: User-driven Intelligent Interface on the Basis of Multimodal Augmented Reality and Brain-Computer Interaction for People with Functional Disabilities. (2017).
<https://doi.org/10.48550/ARXIV.1704.05915>
5. Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A.: Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. IEEE Geosci. Remote Sens. Lett. 14, 778–782 (2017).
<https://doi.org/10.1109/LGRS.2017.2681128>
6. Shelestov, A., Lavreniuk, M., Kussul, N., Novikov, A., Skakun, S.: Exploring Google Earth Engine Platform for Big Data Processing: Classification of Multi-Temporal Satellite Imagery for Crop Mapping. Front. Earth Sci. 5, (2017).
<https://doi.org/10.3389/feart.2017.00017>
7. Kussul, N., Drozd, S., Yailymova, H., Shelestov, A., Lemoine, G., Deininger, K.: Assessing damage to agricultural fields from military actions in Ukraine: An integrated approach using statistical indicators and machine learning. Int. J. Appl.

Earth Obs. Geoinformation. 125, 103562 (2023).
<https://doi.org/10.1016/j.jag.2023.103562>

8. Zhelezniakov, D., Zaytsev, V., Radyvonenko, O.: Acceleration of Online Recognition of 2D Sequences Using Deep Bidirectional LSTM and Dynamic Programming. B: Rojas, I., Joya, G., i Catala, A. (ред.) *Advances in Computational Intelligence*. pp. 438–449. Springer International Publishing, Cham (2019)
9. Lavreniuk, M., Bhat, S.F., Müller, M., Wonka, P.: EVP: Enhanced Visual Perception Using Inverse Multi-attentive Feature Refinement and Regularized Image-Text Alignment. B: Del Bue, A., Canton, C., Pont-Tuset, J., i Tommasi, T. (ред.) *Computer Vision – ECCV 2024 Workshops*. pp. 206–225. Springer Nature Switzerland, Cham (2025)
10. Ramirez, P.Z., Tosi, F., Di Stefano, L., Timofte, R., Costanzino, A., Poggi, M., Salti, S., Mattoccia, S., Zhang, Z., Yang, Y., Chen, W., Ming, A., Zhao, M., Yu, M., Gao, S., Wang, X., Xue, F., Shi, J., Yang, Y., A, Y., Jin, Y., Li, D., Shukla, A., Frija-Altarc, L., Toews, M., Geng, H., Wan, T., Gao, Z., Xu, Q., Xu, K., Zang, Z., Pinjari, J.B., Purohit, K., Lavreniuk, M., Cao, J., Li, S., Jiang, K., Jiang, J., Huang, Y.: NTIRE 2025 Challenge on HR Depth From Images of Specular and Transparent Surfaces. B: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 978–992. IEEE, Nashville, TN, USA (2025)
11. Lavreniuk, M., Lavreniuk, A.: SPIdepth: Strengthened Pose Information for Self-Supervised Monocular Depth Estimation. B: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 865–875. IEEE, Nashville, TN, USA (2025)
12. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *IEEE Trans. Robot.* 31, 1147–1163 (2015).
<https://doi.org/10.1109/TRO.2015.2463671>
13. Chen, J., Chen, J., Liu, K., Chang, H., Fu, S., Yang, J.: Robust 6DoF Pose Tracking Considering Contour and Interior Correspondence Uncertainty for AR

- Assembly Guidance. IEEE Trans. Instrum. Meas. 74, 1–16 (2025).
<https://doi.org/10.1109/TIM.2025.3571175>
14. Murez, Z., Van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: End-to-End 3D Scene Reconstruction from Posed Images. B: Vedaldi, A., Bischof, H., Brox, T., i Frahm, J.-M. (ред.) Computer Vision – ECCV 2020. pp. 414–431. Springer International Publishing, Cham (2020)
 15. Sayed, M., Gibson, J., Watson, J., Prisacariu, V., Firman, M., Godard, C.: SimpleRecon: 3D Reconstruction Without 3D Convolutions. B: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., i Hassner, T. (ред.) Computer Vision – ECCV 2022. pp. 1–19. Springer Nature Switzerland, Cham (2022)
 16. Watson, J., Vicente, S., Aodha, O.M., Godard, C., Brostow, G., Firman, M.: Heightfields for Efficient Scene Reconstruction for AR. B: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5839–5849. IEEE, Waikoloa, HI, USA (2023)
 17. Augmented Reality Market Size, Share, and Growth Analysis, <https://www.skyquestt.com/report/augmented-reality-market>, (2025)
 18. Савін В.В., Аналіз методів 3D реконструкції середовища для доповненої реальності, Методи комп'ютерного зору і глибинних нейронних мереж для еколого-економічного аналізу: монографія / за ред. Н.М. Куссуль, А.Ю. Шелестова – Київ: Наукова думка, 2024. – 448с. С. 49 – 81, ISBN 978-966-00-1940-9.
 19. Bader, M.: Space-filling curves: an introduction with applications in scientific computing. Springer Science & Business Media (2012)
 20. Uss M., Iermolenko R., Kolodiazhna O., Savin V., Method and device for generating depth map, Patent: US20240144503A1, URL: <https://patents.google.com/patent/US20240144503A1>.
 21. Uss M., Iermolenko R., Shashko O., Kolodiazhna O., Safonov I., Savin V., Yeo Y, Ji S., Jeong J., Predicting High-precision Depth on Low-Precision Devices Using 2D Hilbert Curves, Proceedings of the 42nd International Conference on Machine Learning, PMLR, 2025, vol. 267, pp. 60635 – 60656, ISSN: 2640-3498

22. Savin V., Kolodiazhna O., Adapting Neural Radiance Fields (NeRF) to the 3D Scene Reconstruction Problem Under Dynamic Illumination Conditions, *Cybernetics and Systems Analysis*, 2023, vol. 59, pp. 910 – 918, ISSN: 1060-0396, DOI: 10.1007/s10559-023-00626-7
23. Montgomery, D., Peck, E., Vining, G.: *Introduction to Linear Regression Analysis*. John Wiley & Sons (2012)
24. Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* 65, 386–408 (1958). <https://doi.org/10.1037/h0042519>
25. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. *Nature.* 323, 533–536 (1986). <https://doi.org/10.1038/323533a0>
26. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification And Regression Trees*. Routledge (2017)
27. Cortes, C., Vapnik, V.: Support-Vector Networks. *Mach. Learn.* 20, 273–297 (1995). <https://doi.org/10.1023/A:1022627411411>
28. Macqueen, J.: Some methods for classification and analysis of multivariate observations. *Proc. 5-Th Berkeley Symp. Math. Stat. Probab.* Calif. Press. (1967)
29. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature.* 521, 436–444 (2015). <https://doi.org/10.1038/nature14539>
30. Fuhrmann, A., Schmalstieg, D., Purgathofer, W.: *Practical Calibration Procedures for Augmented Reality*, <http://diglib.eg.org/handle/10.2312/EGVE.EGVE00.003-012>, (2000)
31. Cutolo, F., Fontana, U., Cattari, N., Ferrari, V.: Off-Line Camera-Based Calibration for Optical See-Through Head-Mounted Displays. *Appl. Sci.* 10, 193 (2019). <https://doi.org/10.3390/app10010193>
32. Iermolenko R., Sukhariev A., Morozov K., Vdovychenko I., Savin V., Vavdiuk D., Klimenkov O., Sapozhnik O., *Electronic apparatus and controlling method thereof*, Patent: US20230254568A1, URL: <https://patents.google.com/patent/US20230254568A1>.

33. Vdovychenko I., Sapozhnik O., Dykyi V., Savin V., Vitiuk A., Tuzhykov A., Electronic device and method for providing augmented reality environment including adaptive multi-camera, Patent: US20240104867A1, URL: <https://patents.google.com/patent/US20240104867A1>.
34. Omelchenko A., Vdovychenko I., Morozov K., Androsov V., Savin V., Electronic device for controlling audio device on basis of image context, and method for operating same, Patent: US20250193598A1, URL: <https://patents.google.com/patent/US20250193598A1>.
35. Chandan, K., Kudalkar, V., Li, X., Zhang, S.: ARROCH: Augmented Reality for Robots Collaborating with a Human, <https://arxiv.org/abs/2109.10400>, (2021)
36. Schäfer, A., Reis, G., Stricker, D.: AnyGesture: Arbitrary One-Handed Gestures for Augmented, Virtual, and Mixed Reality Applications. Appl. Sci. 12, 1888 (2022). <https://doi.org/10.3390/app12041888>
37. Escobar, M., Puentes, J., Forigua, C., Pont-Tuset, J., Maninis, K.-K., Arbelaez, P.: EgoCast: Forecasting Egocentric Human Pose in the Wild. B: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5831–5841. IEEE, Tucson, AZ, USA (2025)
38. Marchenko A., Savin V., Tymchyshyn V., Gesture sensing method and electronic device supporting same, Patent: US20180329501A1, URL: <https://patents.google.com/patent/US20180329501A1>
39. Sydorenko D., Alkhimova S., Savin V., Shcherbina A., Kim G., Bondarets I., Electronic device and method for identifying relevant device in augmented reality mode of electronic device, Patent: US20220070431A1, URL: <https://patents.google.com/patent/US20220070431A1>
40. Shcherbina A., Bondarets I., Trunov O., Olshevskyi V., Savin V., Method of adaptive 6DoF hand parameters estimation for precise interaction in AR, Patent: US20230004214A1, URL: <https://patents.google.com/patent/US20230004214A1>

41. Saidin, N.F., Abd Halim, N.D., Yahaya, N.: A Review of Research on Augmented Reality in Education: Advantages and Applications. *Int. Educ. Stud.* 8, p1 (2015). <https://doi.org/10.5539/ies.v8n13p1>
42. Савін В. В., Блохіна І. О., Використання технологій доповненої та віртуальної реальності в умовах дистанційного навчання, *Наука та освіта в дослідженнях молодих учених: матеріали IV Міжнар. наук.-практ. конф. для студ., аспірантів, докторантів, молодих учених, Харків, 18 трав. 2023 р. / Харків. нац. пед. ун-т ім. Г. С. Сковороди.*, URL: <https://dspace.hnpu.edu.ua/items/25b98239-a99a-476b-b170-029b6b10f161>
43. Zhu, E., Hadadgar, A., Masiello, I., Zary, N.: Augmented reality in healthcare education: an integrative review. *PeerJ.* 2, e469 (2014). <https://doi.org/10.7717/peerj.469>
44. Eswaran, M., Gulivindala, A.K., Inkulu, A.K., Raju Bahubalendruni, M.V.A.: Augmented reality-based guidance in product assembly and maintenance/repair perspective: A state of the art review on challenges and opportunities. *Expert Syst. Appl.* 213, 118983 (2023). <https://doi.org/10.1016/j.eswa.2022.118983>
45. Pejisa, T., Kantor, J., Benko, H., Ofek, E., Wilson, A.: Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. B: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing.* pp. 1716–1725. ACM, San Francisco California USA (2016)
46. Choi, Y.-W., Yang, H., Im, S., Kim, J., Lee, S., Heo, J.: Pokémon Go: Experiences of Mobile Game in Augmented Reality among Players over 40s. *Korean J. Leisure Recreat. Park.* 45, 39–53 (2021). <https://doi.org/10.26446/kjlrp.2021.3.45.1.39>
47. Grayson, C.: Holographic Waveguides: What You Need To Know To Understand The Smartglasses Market, <https://www.uploadvr.com/waveguides-smartglasses/>, (2017)
48. Samsung's New Transparent MICRO LED Display Blurs the Boundaries Between Content and Reality, <https://news.samsung.com/global/video-ces-2024->

- samsungs-new-transparent-micro-led-display-blurs-the-boundaries-between-content-and-reality, (2024)
49. Meta Quest Pro, https://www.meta.com/quest/quest-pro/?srsltid=AfmBOop99d_6YpfPPd4UWPKxYLNxwbCCVPPG8OEDoM2LoEkNCo0l9b6m
 50. XREAL Air 2 Pro, <https://www.xreal.com/>
 51. Magic Leap 2, <https://www.magicleap.com/magic-leap-2>
 52. What is a ToF sensor?, <https://www.e-consystems.com/blog/camera/technology/what-is-a-time-of-flight-sensor-what-are-the-key-components-of-a-time-of-flight-camera/>, (2021)
 53. What is LiDAR?, <https://www.synopsys.com/glossary/what-is-lidar.html>
 54. What is Structured Light Imaging?, <https://www.roboticstomorrow.com/article/2018/04/what-is-structured-light-imaging/11821>, (2018)
 55. Battiato, S., Curti, S., La Cascia, M., Tortora, M., Scordato, E.: Depth map generation by image classification. Представлена на Electronic Imaging 2004 Квітень 16 (2004)
 56. Cozman, F., Krotkov, E.: Depth from scattering. B: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 801–806. IEEE Comput. Soc, San Juan, Puerto Rico (1997)
 57. Tao, M.W., Srinivasan, P.P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. B: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1940–1948. IEEE, Boston, MA, USA (2015)
 58. Trucco, E., Verri, A.: Introductory techniques for 3-D computer vision. Prentice Hall, Upper Saddle River, NJ (1998)
 59. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. B: Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001). pp. 131–140. IEEE Comput. Soc, Kauai, HI, USA (2001)

60. Stereokinetic Phenomenon from Michael Bach's «Optical Illusions & Visual Phenomena», <https://michaelbach.de/ot/mot-ske/index.html>
61. Parallax scrolling, https://en.wikipedia.org/wiki/Parallax_scrolling
62. Depth of Scene from Depth of Field. IEEE Trans. Pattern Anal. Mach. Intell. 9, 523–531 (1987)
63. Matsuyama, T.: Exploitation of 3D video technologies. B: International Conference on Informatics Research for Development of Knowledge Society Infrastructure, 2004. ICKS 2004. pp. 7–14. IEEE, Kyoto, Japan (2004)
64. Lowe, D.G.: Object recognition from local scale-invariant features. B: Proceedings of the Seventh IEEE International Conference on Computer Vision. pp. 1150–1157 вып.2. IEEE, Kerkyra, Greece (1999)
65. Cui, Z., Tan, P.: Global Structure-from-Motion by Similarity Averaging. B: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 864–872. IEEE, Santiago, Chile (2015)
66. El Hazzat, S., Saaidi, A., Satori, K.: Structure from motion for 3D object reconstruction based on local and global bundle adjustment. B: 2015 Third World Conference on Complex Systems (WCCS). pp. 1–6 (2015)
67. Yin, H., Yu, H.: Incremental SFM 3D reconstruction based on monocular. B: 2020 13th International Symposium on Computational Intelligence and Design (ISCID). pp. 17–21. IEEE, Hangzhou, China (2020)
68. Furukawa, Y., Hernández, C.: Multi-View Stereo: A Tutorial. Found. Trends® Comput. Graph. Vis. 9, 1–148 (2015). <https://doi.org/10.1561/06000000052>
69. Furukawa, Y., Ponce, J.: Accurate, Dense, and Robust Multiview Stereopsis. IEEE Trans. Pattern Anal. Mach. Intell. 32, 1362–1376 (2010). <https://doi.org/10.1109/TPAMI.2009.161>
70. Juszczak, J.M., Wijata, A., Czajkowska, J., Krecichwost, M., Rudzki, M., Biesok, M., Pycinski, B., Majewski, J., Kostecki, J., Pietka, E.: Wound 3D Geometrical Feature Estimation Using Poisson Reconstruction. IEEE Access. 9, 7894–7907 (2021). <https://doi.org/10.1109/ACCESS.2020.3035125>

71. Cai, Y., Cao, M., Li, L., Liu, X.: An End-to-End Approach to Reconstructing 3D Model From Image Set. *IEEE Access*. 8, 193268–193284 (2020). <https://doi.org/10.1109/ACCESS.2020.3032169>
72. Weilharter, R., Fraundorfer, F.: HighRes-MVSNet: A Fast Multi-View Stereo Network for Dense 3D Reconstruction From High-Resolution Images. *IEEE Access*. 9, 11306–11315 (2021). <https://doi.org/10.1109/ACCESS.2021.3050556>
73. Chen, P.-H., Yang, H.-C., Chen, K.-W., Chen, Y.-S.: MVSNet++: Learning Depth-Based Attention Pyramid Features for Multi-View Stereo. *IEEE Trans. Image Process*. 29, 7261–7273 (2020). <https://doi.org/10.1109/TIP.2020.3000611>
74. Lu, Q., Lu, Y., Xiao, M., Yuan, X., Jia, W.: 3D-FHNet: Three-Dimensional Fusion Hierarchical Reconstruction Method for Any Number of Views. *IEEE Access*. 7, 172902–172912 (2019). <https://doi.org/10.1109/ACCESS.2019.2955288>
75. Tan, M., Le, Q.V.: EfficientNetV2: Smaller Models and Faster Training. (2021). <https://doi.org/10.48550/ARXIV.2104.00298>
76. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. B: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (2016)
77. Duzceker, A., Galliani, S., Vogel, C., Speciale, P., Dusmanu, M., Pollefeys, M.: DeepVideoMVS: Multi-View Stereo on Video with Recurrent Spatio-Temporal Fusion. B: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15319–15328. IEEE, Nashville, TN, USA (2021)
78. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. B: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9492–9502. IEEE, Seattle, WA, USA (2024)
79. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. B: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685. IEEE, New Orleans, LA, USA (2022)

80. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5B: An open large-scale dataset for training next generation image-text models, <https://arxiv.org/abs/2210.08402>, (2022)
81. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. B: Vedaldi, A., Bischof, H., Brox, T., i Frahm, J.-M. (ред.) Computer Vision – ECCV 2020. pp. 405–421. Springer International Publishing, Cham (2020)
82. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH Comput. Graph. 18, 165–174 (1984). <https://doi.org/10.1145/964965.808594>
83. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D Gaussian Splatting for Real-Time Radiance Field Rendering. (2023). <https://doi.org/10.48550/ARXIV.2308.04079>
84. Waechter, M., Moehrle, N., Goesele, M.: Let There Be Color! Large-Scale Texturing of 3D Reconstructions. B: Fleet, D., Pajdla, T., Schiele, B., i Tuytelaars, T. (ред.) Computer Vision – ECCV 2014. pp. 836–850. Springer International Publishing, Cham (2014)
85. Jiang, C., Sud, A., Makadia, A., Huang, J., NieBner, M., Funkhouser, T.: Local Implicit Grid Representations for 3D Scenes. B: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6000–6009. IEEE, Seattle, WA, USA (2020)
86. Penner, E., Zhang, L.: Soft 3D reconstruction for view synthesis. ACM Trans. Graph. 36, 1–11 (2017). <https://doi.org/10.1145/3130800.3130855>
87. Lin, C.-H., Ma, W.-C., Torralba, A., Lucey, S.: BARF: Bundle-Adjusting Neural Radiance Fields. B: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5721–5731. IEEE, Montreal, QC, Canada (2021)
88. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. B: 2021

- IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14104–14113. IEEE, Montreal, QC, Canada (2021)
89. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo. B: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5590–5599. IEEE, Montreal, QC, Canada (2021)
 90. Schonberger, J.L., Frahm, J.-M.: Structure-from-Motion Revisited. B: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113. IEEE, Las Vegas, NV, USA (2016)
 91. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Niebner, M.: Dense Depth Priors for Neural Radiance Fields from Sparse Input Views. B: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12882–12891. IEEE, New Orleans, LA, USA (2022)
 92. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. B: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6494–6504. IEEE, Nashville, TN, USA (2021)
 93. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. B: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10892–10902. IEEE, Montreal, QC, Canada (2021)
 94. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. B: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
 95. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Smagt, P.V.D., Cremers, D., Brox, T.: FlowNet: Learning Optical Flow with Convolutional Networks. B: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2758–2766. IEEE, Santiago (2015)

96. Guo, X., Zhang, C., Zhang, Y., Zheng, W., Nie, D., Poggi, M., Chen, L.: Lightstereo: Channel Boost is All You Need for Efficient 2D Cost Aggregation. B: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 8738–8744. IEEE, Atlanta, GA, USA (2025)
97. Shamsafar, F., Woerz, S., Rahim, R., Zell, A.: MobileStereoNet: Towards Lightweight Deep Networks for Stereo Matching. B: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 677–686. IEEE, Waikoloa, HI, USA (2022)
98. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks. B: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4510–4520. IEEE, Salt Lake City, UT (2018)
99. Wang, Y., Lai, Z., Huang, G., Wang, B.H., van der Maaten, L., Campbell, M., Weinberger, K.Q.: Anytime Stereo Image Depth Estimation on Mobile Devices, <https://arxiv.org/abs/1810.11408>, (2018)
100. Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.-H., Kautz, J.: Learning Affinity via Spatial Propagation Networks. B: Advances in neural information processing systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017). pp. 1519–1529. Curran Associates Inc.57 Morehouse LaneRed HookNYUnited States, Long Beach California USA (2017)
101. Jiang, X., Cambareri, V., Agresti, G., Ugwu, C.I., Simonetto, A., Cardinaux, F., Zanuttigh, P.: A Low Memory Footprint Quantized Neural Network for Depth Completion of Very Sparse Time-of-Flight Depth Maps. B: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 2686–2695. IEEE, New Orleans, LA, USA (2022)
102. Chen, Z., Badrinarayanan, V., Drozdov, G., Rabinovich, A.: Estimating Depth from RGB and Sparse Sensing. (2018). <https://doi.org/10.48550/ARXIV.1804.02771>
103. Mehta, S., Rastegari, M.: Separable Self-attention for Mobile Vision Transformers, <https://arxiv.org/abs/2206.02680>, (2022)

104. Mehta, S., Rastegari, M.: MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer, <https://arxiv.org/abs/2110.02178>, (2021)
105. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, <https://arxiv.org/abs/1409.0575>, (2014)
106. Laga, H., Jospin, L.V., Boussaid, F., Bennamoun, M.: A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 1738–1764 (2022). <https://doi.org/10.1109/TPAMI.2020.3032602>
107. Watson, J., Aodha, O.M., Turmukhambetov, D., Brostow, G.J., Firman, M.: Learning Stereo from Single Images. B: Vedaldi, A., Bischof, H., Brox, T., i Frahm, J.-M. (ред.) *Computer Vision – ECCV 2020*. pp. 722–740. Springer International Publishing, Cham (2020)
108. Moreau, A., Piasco, N., Tsishkou, D., Stanciulescu, B., de La Fortelle, A.: LENS: Localization enhanced by NeRF synthesis, <https://arxiv.org/abs/2110.06558>, (2021)
109. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised Monocular Depth Estimation with Left-Right Consistency. B: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6602–6611. IEEE, Honolulu, HI (2017)
110. Li, H., Gordon, A., Zhao, H., Casser, V., Angelova, A.: Unsupervised Monocular Depth Learning in Dynamic Scenes, <https://arxiv.org/abs/2010.16404>, (2020)
111. Wang, K., Zhang, Z., Yan, Z., Li, X., Xu, B., Li, J., Yang, J.: Regularizing Nighttime Weirdness: Efficient Self-supervised Monocular Depth Estimation in the Dark. B: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16035–16044. IEEE, Montreal, QC, Canada (2021)
112. Kolodiazhna O., Savin V., Uss M., Kussul N., 3D Scene Reconstruction with Neural Radiance Fields (NeRF) considering dynamic illumination conditions, *Proceedings of International Conference on Applied Innovation in IT 2023*, 2023, Volume 11, Issue 1, pp. 233-238. ISSN: 2199-8876, DOI: 10.25673/101943

113. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. B: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
114. Kusupati, U., Cheng, S., Chen, R., Su, H.: Normal Assisted Stereo Depth Estimation. B: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2186–2196. IEEE, Seattle, WA, USA (2020)
115. Wonhyeok, C., Kyumin, H., Wei, P., Minwoo, C., Sunghoon, I.: Self-supervised Monocular Depth Estimation Robust to Reflective Surface Leveraged by Triplet Mining. Представлена на International Conference on Learning Representations (ICLR 2025) КВІТЕНЬ 24 (2025)
116. Yao, J., Wu, T., Zhang, X.: Improving Depth Gradient Continuity in Transformers: A Comparative Study on Monocular Depth Estimation with CNN, <https://arxiv.org/abs/2308.08333>, (2023)
117. Li, Y., Wei, X.: MobileDepth: Monocular Depth Estimation Based on Lightweight Vision Transformer. Appl. Artif. Intell. 38, 2364159 (2024). <https://doi.org/10.1080/08839514.2024.2364159>
118. Henley, C., Somasundaram, S., Hollmann, J., Raskar, R.: Detection and mapping of specular surfaces using multibounce LiDAR returns. Opt. Express. 31, 6370 (2023). <https://doi.org/10.1364/OE.479900>
119. Feng, W., Cheng, X., Sun, J., Xiong, Z., Zhai, Z.: Specular highlight removal and depth estimation based on polarization characteristics of light field. Opt. Commun. 537, 129467 (2023). <https://doi.org/10.1016/j.optcom.2023.129467>
120. Bhat, D.N., Nayar, S.K.: Stereo in the presence of specular reflection. B: Proceedings of IEEE International Conference on Computer Vision. pp. 1086–1092. IEEE Comput. Soc. Press, Cambridge, MA, USA (1995)
121. Yeshwanth, C., Liu, Y.-C., Nießner, M., Dai, A.: ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes. B: Int. Conf. Comput. Vis. pp. 12–22 (2023)
122. Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object

- Detection for Autonomous Driving. B: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8437–8445 (2019)
123. You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M.E., Weinberger, K.Q.: Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving. B: Int. Conf. Learn. Represent. (2020)
 124. Wofk, D., Ma, F., Yang, T.-J., Karaman, S., Sze, V.: FastDepth: Fast Monocular Depth Estimation on Embedded Systems. B: Int. Conf. on Robotics and Automation (ICRA). pp. 6101–6108 (2019)
 125. Rasla, A., Beyeler, M.: The Relative Importance of Depth Cues and Semantic Edges for Indoor Mobility Using Simulated Prosthetic Vision in Immersive Virtual Reality. B: Proc. of ACM Symp. on Virtual Reality Software and Tech. Association for Computing Machinery, New York, NY, USA (2022)
 126. Ignatov, A., Malivenko, G., Timofte, R., Treszczotko, L., Chang, X., Ksiazek, P., Lopuszynski, M., Pioro, M., Rudnicki, R., Smyl, M., Ma, Y., Li, Z., Chen, Z., Xu, J., Liu, X., Jiang, J., Shi, X., Xu, D., Li, Y., Wang, X., Lei, L., Zhang, Z., Wang, Y., Huang, Z., Luo, G., Yu, G., Fu, B., Li, J., Wang, Y., Huang, Z., Cao, Z., Conde, M.V., Sapozhnikov, D., Lee, B.H., Park, D., Hong, S., Lee, J., Lee, S., Chun, S.Y.: Efficient Single-Image Depth Estimation on Mobile Devices, Mobile AI & AIM 2022 Challenge: Report. B: Eur. Conf. Comput. Vis. pp. 71–91 (2022)
 127. Kolbeinsson, B., Mikolajczyk, K.: DDOS: The Drone Depth and Obstacle Segmentation Dataset. B: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 7328–7337 (2024)
 128. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: Channel-Aware Distillation Transformer for Depth Estimation on Nano Drones. arXiv. (2023). <https://doi.org/10.48550/arXiv.2303.10386>
 129. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. B: Int. Conf. Comput. Vis. pp. 12179–12188 (2021)
 130. Bao, H., Dong, L., Wei, F.: BEiT: BERT Pre-Training of Image Transformers. B: Int. Conf. Learn. Represent. (2022)

131. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.Q., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y. (Bernie), Li, S.-W., Misra, I., Rabbat, M.G., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning Robust Visual Features without Supervision. *Trans Mach. Learn. Res.* (2024). <https://doi.org/10.48550/arXiv.2304.07193>
132. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. *IEEE Conf Comput Vis Pattern Recog.* 10371–10381 (2024). <https://doi.org/10.1109/CVPR52733.2024.00987>
133. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. *arXiv.* (2024). <https://doi.org/10.48550/arXiv.2406.09414>
134. Shao, J., Yang, Y., Zhou, H., Zhang, Y., Shen, Y., Poggi, M., Liao, Y.: Learning Temporally Consistent Video Depth from Video Diffusion Priors. *arXiv.* (2024). <https://doi.org/10.48550/arXiv.2406.01493>
135. Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching. *B: Int. Conf. on 3D Vision (3DV).* pp. 218–227 (2021)
136. Xianqi, W., Gangwei, X., Hao, J., Xin, Y.: Selective-Stereo: Adaptive Frequency Information Selection for Stereo Matching. *B: IEEE Conf. Comput. Vis. Pattern Recog.* pp. 19701–19710 (2024)
137. Nicolas, B., Li, Y., Chris, P., Aaron, C.: Delving Deeper into Convolutional Networks for Learning Video Representations. *arXiv.* (2015). <https://doi.org/10.48550/arXiv.1511.06432>
138. Oh, S., Kim, H.-J.S., Lee, J., Kim, J.: RRNet: Repetition-reduction network for energy efficient depth estimation. *IEEE Access.* 8, 106097–106108 (2020). <https://doi.org/10.1109/ACCESS.2020.3000773>
139. Liu, Q., Zhou, S.: LightDepthNet: Lightweight CNN architecture for monocular depth estimation on edge devices. *IEEE Trans Circuits Syst. II Express Briefs.* 71, 2389–2393 (2023). <https://doi.org/10.1109/TCSII.2023.3337369>

140. Song, L., Shi, D., Xia, J., Ouyang, Q., Qiao, Z., Jin, S., Yang, S.: Spatial-aware dynamic lightweight self-supervised monocular depth estimation. *IEEE Robot. Autom. Lett.* 9, 883–890 (2023). <https://doi.org/10.1109/LRA.2023.3337991>
141. Li, G., Xue, J., Liu, L., Wang, X., Ma, X., Dong, X., Li, J., Feng, X.: Unleashing the Low-Precision Computation Potential of Tensor Cores on GPUs. B: *IEEE/ACM Int. Symp. on Code Generation and Optimization (CGO)*. pp. 90–102 (2021)
142. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., Kalenichenko, D.: Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. B: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2704–2713 (2018)
143. Nagel, M., Fournarakis, M., Amjad, R.A., Bondarenko, Y., Baalen, M.V., Blankevoort, T.: A white paper on neural network quantization. *arXiv*. (2021). <https://doi.org/10.48550/arXiv.2106.08295>
144. Yang, J., Shen, X., Xing, J., Tian, X., Li, H., Deng, B., Huang, J., Hua, X.: Quantization networks. B: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 7308–7316 (2019)
145. Sakr, C., Dai, S., Venkatesan, R., Zimmer, B., Dally, W., Khailany, B.: Optimal clipping and magnitude-aware differentiation for improved quantization-aware training. B: *Int. Conf. Mach. Learn.* pp. 19123–19138. PMLR (2022)
146. Bhalgat, Y., Lee, J., Nagel, M., Blankevoort, T., Kwak, N.: LSQ+: Improving low-bit quantization through learnable offsets and better initialization. B: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 2978–2985 (2020)
147. Li, Z., Yang, T., Wang, P., Cheng, J.: Q-ViT: Fully differentiable quantization for vision transformer. *arXiv*. (2022). <https://doi.org/10.48550/arXiv.2201.07703>
148. Fang, J., Shafiee, A., Abdel-Aziz, H., Thorsley, D., Georgiadis, G., Hassoun, J.H.: Post-training piecewise linear quantization for deep neural networks. B: *Eur. Conf. Comput. Vis.* pp. 69–86. Springer (2020)
149. Sheng, T., Feng, C., Zhuo, S., Zhang, X., Shen, L., Aleksic, M.: A quantization friendly separable convolution for MobileNets. B: *1st Workshop on Energy*

- Efficient Machine Learning and Cognitive Computing for Embedded Applications (EMC2). pp. 14–18 (2018)
150. Wang, T., Wang, K., Cai, H., Lin, J., Liu, Z., Han, S.: APQ: Joint Search for Network Architecture, Pruning and Quantization Policy. B: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
 151. Shlezinger, N., Eldar, Y.C., Rodrigues, M.R.D.: Hardware-Limited Task-Based Quantization. B: IEEE Int. Workshop on Signal Process. Advances in Wireless Communications (SPAWC). pp. 1–5 (2019)
 152. Kiyama, M., Nakahara, Y., Amagasaki, M., Iida, M.: A Quantized Neural Network Library for Proper Implementation of Hardware Emulation. B: Int. Symp. on Computing and Networking Workshops (CANDARW). pp. 136–140 (2019)
 153. Wang, K., Liu, Z., Lin, Y., Lin, J., Han, S.: HAQ: Hardware-Aware Automated Quantization With Mixed Precision. B: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8604–8612 (2019)
 154. Borrego-Carazo, J., Ozay, M., Laboyrie, F., Wisbey, P.: A Mixed Quantization Network for Computationally Efficient Mobile Inverse Tone Mapping. Br. Mach. Vis. Conf. (2021). <https://doi.org/10.48550/arXiv.2203.06504>
 155. Qualcomm Technologies, Inc.: Snapdragon neural processing engine SDK, <https://www.qualcomm.com/developer/software/neural-processing-sdk-for-ai>, (2025)
 156. Google: Coral Edge TPU Documentation, <https://coral.ai/docs/edgetpu/models-intro/#model-requirements>, (2020)
 157. Jiang, X., Cambareri, V., Agresti, G., Ugwu, C.I., Simonetto, A., Cardinaux, F., Zanuttigh, P.: A Low Memory Footprint Quantized Neural Network for Depth Completion of Very Sparse Time-of-Flight Depth Maps. B: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 2687–2696 (2022)
 158. Sagan, H.: Space-filling curves. Springer Science & Business Media (2012)
 159. Ventrella, J.: Brainfilling curves – a fractal bestiary. Eyebrian Books (2012)

160. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries. B: IEEE Winter Conf. on Applications of Comput. Vis. (WACV). pp. 1043–1051 (2019)
161. Wadekar, S.N., Chaurasia, A.: MobileViTv3: mobile-friendly vision transformer with simple and effective fusion of local, global and input features. arXiv. (2022). <https://doi.org/10.48550/arXiv.2209.15159>
162. Shkolnik, M., Chmiel, B., Banner, R., Shomron, G., Nahshan, Y., Bronstein, A., Weiser, U.: Robust quantization: one model to rule them all. B: Adv. Neural Inform. Process. Syst. (2020)
163. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Adv Neural Inf. Process Syst. 2366–2374 (2014). <https://doi.org/10.5555/2969033.2969091>
164. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. B: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3354–3361 (2012)
165. Wang, Z., Bovik, A.C., Sheikh, H.R.: Structural similarity based image quality assessment. Digit. Video Image Qual. Percept. Coding Ser Ser. Signal Process Commun. (2005). <https://doi.org/10.1201/9781420027822.ch7>
166. Rouse, D.M., Hemami, S.S.: Understanding and simplifying the structural similarity metric. B: IEEE Int. Conf. Image Process. pp. 1188–1191 (2008)
167. Lahitani, A.R., Permanasari, A.E., Setiawan, N.A.: Cosine similarity to determine similarity measure: study case in online essay assessment. B: Int. Conf. on Cyber and IT Service Management. pp. 1–6 (2016)
168. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE Trans Comput. 23, 90–93 (1974). <https://doi.org/10.1109/T-C.1974.223784>
169. Rousseeuw, P.J., Croux, C.: Alternatives to the Median Absolute Deviation. J Am. Stat. Assoc. 88, 1273–1283 (1993). <https://doi.org/10.1080/01621459.1993.10476408>

170. Cabon, Y., Murray, N., Humenberger, M.: Virtual KITTI 2. arXiv. (2020). <https://doi.org/10.48550/arXiv.2001.10773>
171. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. B: Proc. of Annual Conf. on Computer Graphics and Interactive Techniques. pp. 303-- 312. Association for Computing Machinery, New York, NY, USA (1996)
172. Zhou, Q.-Y., Park, J., Koltun, V.: Open3D: A Modern Library for 3D Data Processing. arXiv. (2018). <https://doi.org/10.48550/arXiv.1801.09847>
173. Toshev, A., Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Networks. B: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1653–1660 (2014)
174. Xiao, B., Wu, H., Wei, Y.: Simple Baselines for Human Pose Estimation and Tracking. B: Eur. Conf. Comput. Vis. pp. 472–487 (2018)
175. Lin, T.-Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. B: Eur. Conf. Comput. Vis. (2014)
176. Bello, I., Fedus, W., Du, X., Cubuk, E.D., Srinivas, A., Lin, T.-Y., Shlens, J., Zoph, B.: Revisiting ResNets: Improved Training and Scaling Strategies. B: Int. Conf. on Neural Information Processing Systems. pp. 22614–2262. Curran Associates Inc., Red Hook, NY, USA (2021)
177. Xu, G., Wang, X., Ding, X., Yang, X.: Iterative geometry encoding volume for stereo matching. B: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21919–21928 (2023)
178. Shao, S., Pei, Z., Chen, W., Wu, X., Li, Z.: NDDepth: Normal-distance assisted monocular depth estimation. B: Int. Conf. Comput. Vis. pp. 7931–7940 (2023)

ЗАТВЕРДЖЕНО

Директор ТОВ
«Самсунг РнД Інститут Україна»
к.ф.-м.н. Фісуненко Андрій Леонідович



«19» листопада 2025р.

АКТ

впровадження результатів досліджень дисертаційної роботи
Савіна Володимира Вадимовича, аспіранта кафедри математичного моделювання та аналізу
даних Навчально-наукового Фізико-технічного інституту Національного технічного
університету України «Київський політехнічний інститут імені Ігоря Сікорського», на тему
«Математичні моделі та методи 3D реконструкції для доповненої реальності», яка готується
до подання на захист для здобуття наукового ступеня Доктора філософії за спеціальністю 113
«Прикладна математика»

Підрозділ досліджень та розробки у галузі візуального інтелекту компанії «Самсунг
РнД Інститут Україна» повідомляє, що, у контексті вирішення завдань 3D реконструкції сцени
для доповненої реальності та створення/редагування просторового контенту, було використано
результати дисертаційної роботи Савіна В.В. за темою «Математичні моделі та методи 3D
реконструкції для доповненої реальності», що базуються на відкритих публікаціях у світових
та українських виданнях та конференціях.

Метод призначений для 3D реконструкції карт глибини з урахуванням напівпрозорих та
відбивних поверхонь, що орієнтований на підвищення якості відтворення складних об'єктів,
зокрема при наявності динамічного освітлення захищено патентом (US20240144503A1)
Отриманий патент розширює патентне портфоліо компанії у відповідному технологічному
домені.

Метод прогнозування високоточних карт глибини з широким діапазоном на пристроях
з обмеженою розрядністю, що ґрунтується на використанні двовимірних кривих Гільберта,
дозволяє: зменшити похибку квантування моделі у 4,6 рази, що підвищує якість реконструкції
карт глибини на DSP-пристроях; розширити діапазон відстаней за рахунок збільшення
ефективної розрядності з 8-біт до 10-біт; зменшити час виконання та енергоспоживання
квантованої моделі у 1,5 раза порівняно з оригінальною моделлю за умови збереження або
покращення якості прогнозування карт глибини.

Розглянуті та запропоновані Савіним В.В методи 3D реконструкції сцени на кінцевих
пристроях користувачів з обмеженими обчислювальними ресурсами було використано при
розробці комерційних проєктів, що спрямовані на сценарії створення/редагування
просторового контенту для флагманської моделі смартфона Samsung Galaxy S25.

УЗГОДЖЕНО

Заступник директора з технологій штучного
інтелекту
«Самсунг РнД Інститут Україна»

«19» листопада 2025
С. Ю. Литвиненко