НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
Навчально-науковий фізико-технічний інститут
Кафедра математичного моделювання та аналізу даних

«На правах рукопису»
УДК 004.93

До захисту допущено:
Завідувач кафедри

_____ Наталія КУССУЛЬ
“_____” _____ 2022 р.

Магістерська дисертація

на здобуття ступеня магістра
за освітньо-професійною програмою «Математичні методи моделювання, розпізнавання образів та комп'ютерного зору»
спеціальності: 113 «Прикладна математика»
на тему: Генерування тренувальних даних за допомогою технології NeRF для задач стереозору

Виконала: студентка II курсу, групи ФІ-11мп
Колодяжна Олена Олександрівна _____

Науковий керівник:
професор, д.т.н., Куссуль Наталія Миколаївна _____

Консультант:
доцент, к.т.н., Усс Михайло Леонтійович _____

Рецензент:
доцент, к.ф.-м.н., Єрмоленко Руслан Вікторович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ — 2022

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені Ігоря СІКОРСЬКОГО»
Навчально-науковий фізико-технічний інститут
Кафедра математичного моделювання та аналізу даних

Рівень вищої освіти — другий (магістерський)
Спеціальність — 113 Прикладна математика,
Освітньо-професійна програма «Математичні методи моделювання, розпізнавання образів та комп'ютерного зору»

ЗАТВЕРДЖУЮ
завідувача кафедри
_____ Наталія КУССУЛЬ
«__» _____ 2022 р.

ЗАВДАННЯ
на магістреську дисертацію студенту

Колодяжна Олена Олександрівна

1. Тема дисертації «Генерування тренувальних даних за допомогою технології NeRF для задач стереозору», науковий керівник дисертації Куссуль Наталія Миколаївна, професор, д.т.н., затверджені наказом по університету №__ від «__» _____ 2022р.

2. Термін подання студентом дисертації: «__» _____ 2022р.

3. Об'єкт дослідження: монокулярні набори даних для оцінки карт глибин, реальні дані

4. Предмет дослідження: генерування стереозображень для задач стереозору за допомогою NeRF

5. Перелік завдань, які потрібно розробити:

1) Вивчити методи генерування стереозображень

2) Опрацювати літературу щодо генерування зображень за допомогою NeRF

3) Розробити математичні моделі, які доповнять обрану модель NeRF

4) Розробити програмне забезпечення

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

презентація доповіді

7. Орієнтовний перелік публікацій: одна публікація

8. Консультанти розділів дисертації:

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата
1	Усс М. Л., головний інженер-програміст Самсунг РнД Інститут Україна	
2	Усс М. Л., головний інженер-програміст Самсунг РнД Інститут Україна	
3	Усс М. Л., головний інженер-програміст Самсунг РнД Інститут Україна	

9. Дата видачі завдання: «19» жовтня 2022р.

Календарний план

№ з/п	Назва етапів виконання магістреської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Узгодження теми роботи із науковим керівником	Вересень-жовтень 2021 р.	Виконано
2	Огляд опублікованих джерел за тематикою дослідження	Жовтень-грудень 2021 р.	Виконано
3	Розробка плану роботи	Грудень-січень 2021-2022 р.	Виконано
4	Пошук та налаштування прототипу моделі для дослідження	Січень-березень 2022 р.	Виконано
5	Розробка модифікацій моделі, проведення експериментів	Квітень-вересень 2022 р.	Виконано
6	Аналіз та оцінка отриманих результатів роботи	Вересень 2022 р.	Виконано
7	Оформлення магістрерської дисертації	Вересень-листопад 2022 р.	Виконано
8	Отримання допуску до захисту	5 грудня 2022 р.	Виконано
9	Захист магістерської дисертації	19 грудня 2022 р.	Виконано

Студент _____ Олена КОЛОДЯЖНА

Науковий керівник _____ Наталія КУССУЛЬ

РЕФЕРАТ

Кваліфікаційна робота містить 75 сторінок, 45 ілюстрацій, 4 таблиці, 57 джерел літератури.

Задача тривимірної реконструкції сцени є однією з центральних в областях комп'ютерного зору та комп'ютерної графіки та має багато застосувань. У останні роки було розроблено чимало різних методів вирішення даної проблеми, серед яких є використання глибоких нейронних мереж. Складність даної задачі полягає в потребі в одночасній узгодженості локальних деталей та глобальних структур, в великих обчисленнях, а також в обсягах даних. Остання проблема відіграє важливу роль при навчанні глибоких нейромереж.

Дана магістерська дисертація досліджує можливості генерування навчальних даних за допомогою Neural Radiance Fields для задач стереозору. Метою роботи, окрім генерування даних, є аналіз їх ефективності у застосуванні для навчання глибоких нейронних мереж для оцінки карт глибин зі стереозображень.

У результаті було запропоновано певні модифікації моделі BARF, які дають змогу покращити результат синтезу даних, порівняно з оригінальною моделлю, а також запропоновано змінений ланцюжок стандартної підготовки даних та тренування стереонейромереж, який потенційно може дозволити замінити навчання без учителя навчанням з учителем.

Ключові слова: комп'ютерний зір, глибокі нейронні мережі, синтез даних, 3D реконструкція сцени, neural radiance fields, стереобачення

ABSTRACT

The thesis consists of 75 pages, 45 figures, 4 tables, 57 names of bibliographic sources.

3D scene reconstruction is a long-standing problem and a central one in computer vision and computer graphics with many applications. There are many different methods of solving this problem including deep neural networks. The complexity of this task lies in the need for simultaneous consistency of local details and global structures, in large calculations, as well as in the need for a large amount of data. The last problem plays a crucial role in deep neural networks training.

This master's thesis explores the possibilities of generating training data using Neural Radiance Fields for stereo vision problems. The purpose of the work, in addition to data generation, is to analyze their effectiveness in the application for training depth from stereo neural networks.

As a result, certain modifications of the BARF model were proposed, which allow to improve the result of data synthesis compared to the original model, and a modified chain of standard data preparation and training of Depth From Stereo networks was proposed, which could potentially replace unsupervised training with supervised one.

Key words: computer vision, deep neural networks, view synthesis, 3D scene reconstruction, neural radiance fields, stereo vision

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

NeRF - Neural Radiance Fields

AR - доповнена реальність

VR - віртуальна реальність

NN - нейромережа

GT - істинні дані

SDF (Signed Distance Functions) - функції відстаней зі знаком

MLP - багатошаровий перцептрон

BARF - Bundle-Adjusting Neural Radiance Fields

MVS (Multi-View Stereo) - це режим реконструкцій сцени, використовуючи багатовидому геометрію

SLAM (Simultaneous Localization And Mapping) - одночасна локалізація і побудова карти

SfM (Structure from Motion) - метод побудови структури з руху

SSIM (structure Similarity Index) - індекс структурної подібності

CNN - конволюційна нейронна мережа

MSE - середня квадратична помилка

PSNR (Peak Signal-to-Noise Ratio) - пікове співвідношення сигналу до шуму

RMSE - квадратний корінь середньої квадратичної помилки

MAE - середня абсолютна помилка

MARE - середня відносна абсолютна помилка

ВСТУП

Одними з популярних тем у комп'ютерному зорі є AR, VR, autonomous driving. Усі зазначені напрямки досліджень використовують методи стереоспівставлення. Серед задач AR є, наприклад, задача реконструкції тривимірної сцени, що є однією із центральних задач комп'ютерного зору з багатьма застосуваннями. У AR, щоб забезпечити реалістичну та захоплюючу взаємодію між ефектами доповненої реальності та навколишнім реальним світом, 3D-реконструкція має бути точною, узгодженою та виконуватись у реальному часі. А для того, щоб побудувати модель, яка б задовольняла усі перелічені вимоги, потрібен великий обсяг даних. Тому задача генерування даних, які б допомогли вирішити задачі у даних сферах, набуває все більшого інтересу.

Актуальність дослідження

Сучасні методи, які вирішують задачу стереоспівставлення, часто використовують методи глибинного навчання. Однією з проблем є те, що такі глибокі нейромережі потребують великих обсягів даних. Також для задач стереоспівставлення можуть бути потрібні дані з конкретними параметрами камер. Збір датасетів великих об'ємів є важким та затратним, а наявних наборів даних вкрай мало для вирішення поставлених задач. Також проблемою може слугувати те, що потрібні дані покривають тільки певну область, наприклад, дороги та автомобілі. Отже, постає важливе питання в додаткових способах генерування датасетів будь-якої тематики, які б містили у собі стереозображення з відповідними значеннями відстаней до об'єктів.

Мета і завдання дослідження

У даній магістерській дисертації досліджуються NeRF як спосіб генерування зображень та методи його покращення. Метою роботи є дослідження можливості синтезу даних для задач стереозору та їх ефективність у застосуванні для навчання глибоких стереонейромереж

для оцінки карт глибин.

Завдання дослідження:

1) Вивчити методи генерування стерео зображень

2) Опрацювати літературу щодо генерування зображень за допомогою NeRF

3) Розробити математичні моделі, які доповнять обрану модель NeRF

4) Розробити програмне забезпечення для генерування стереонаборів даних, яке використоє отримані математичні моделі

Об'єкт дослідження – монокулярні набори даних для оцінки карт глибин, реальні дані

Предмет дослідження – генерування стереозображень для задач стереозору за допомогою NeRF

Наукова новизна одержаних результатів

Запропоновано використання технології NeRF для генерування/аугментації даних для задачі стереоспівставлення. Запропоновані модифікації моделі NeRF, які покращують якість синтезованих пар за рахунок використання неповних карт глибин та параметру часу.

Практичне значення одержаних результатів

Отримані результати дають змогу синтезувати стереозображення з наявних монокулярних наборів даних та застосовувати їх для покращення якості результатів навчання нейромереж для оцінки карт глибин зі стереозображень.

Апробація результатів роботи

Роботу було оприлюднено та захищено у формі доповіді на XX Всеукраїнській науково-практичній конференції студентів, аспірантів та молодих вчених "Теоретичні і прикладні проблеми фізики, математики та інформатики".

Публікації

Роботу було опубліковано в збірнику матеріалів XX Всеукраїнської науково-практичної конференції студентів, аспірантів та молодих вчених

"Теоретичні і прикладні проблеми фізики, математики та інформатики".

1 ОГЛЯД ІСНУЮЧИХ ПІДХОДІВ ГЕНЕРУВАННЯ ЗОБРАЖЕНЬ ТА ОСНОВНІ ПОНЯТТЯ

Стереоспівставлення є однією з основних технологій комп'ютерного зору, яка спрямована на відновлення 3D-структур реального світу з 2D зображень. Ми можемо оцінити 3D структуру сцени чи об'єкту, знайшовши карти зсувів між ректифікованими лівими та правими зображеннями даної локації [1]. Існують різні підходи до вирішення цієї проблеми, включаючи системи, які базуються на марковських випадкових полях [2], алгоритмах локального стереоспівставлення [3, 4], а також нейронних мережах (NN) [5]. Глибокі нейромережі для оцінки карт глибин зі стереозображень (глибокі стереонейромережі) є одним із найкращих методів досягнення хорошої відповідності між зображеннями стереопари. Щоб навчити нейронну мережу оцінювати глибину, потрібна велика кількість даних. Хоча існує багато доступних монокулярних наборів даних для оцінки глибини, кількість наборів стереоданих для навчання глибоких нейронних мереж є обмеженою. Отже, проблема полягає в тому, що зачасту ми не маємо у своєму розпорядженні наборів даних достатнього об'єму, які містять стереопари з істинними картами глибин (GT) для лівого та правого зображень, і зібрати таку велику кількість даних є складним завданням. Якщо стереомережа розроблюється для роботи з камерою, яка має певні задані параметри, навчальний набір даних із такими ж або подібними параметрами може бути недоступним.

1.1 Задача стереозору

Задача стереозору або оцінка карти зсувів – це процес знаходження пікселів на двох зображеннях, які відповідають одній і тій же точці у 3D

просторі [6]. Мета даного співставлення – це відновлення 3D структури сцени за набором 2D зображень. Приклад стереосистеми зображений на рисунку 1.1. На даному рисунку O_l, O_r – дві камери, які розташовані на одній горизонтальній лінії, I_l, I_r - ліва та права площини зображень, для яких епіполярні лінії (e) є паралельними осі x. Дані площини відповідають тому, що бачать відповідні ліва та права камери, а точка P відповідає проекції двох точок P_l та P_r у 3D.

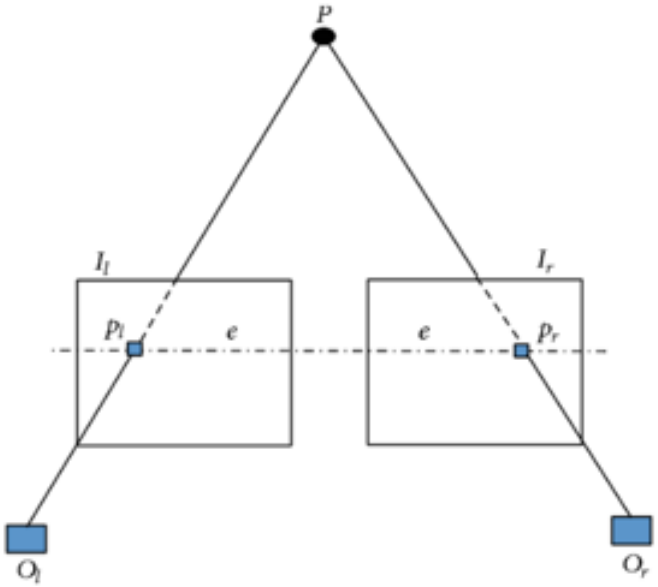


Рисунок 1.1 – Приклад стереосистеми

1.2 Існуючі набори даних та генерування зображень

Для тренування глибоких нейронних мереж для монокулярної оцінки карт глибин потрібно мати зображення та відповідні істинні карти відстаней, а для оцінки зі стереопар потрібно відповідно стереозображення та відповідні карти глибин. Також зазначені нейромережі вимагають великих обсягів даних, що досягають десятки тисяч зображень та більше. Наборів даних зазначеної кількості у відкритому доступі досить мало. Прикладами стереодатасетів є, наприклад, реальні дані для задач, пов'язаних з автомобілями та

дорогами, KITTY [7], DrivingStereo [8]. Також до наявних наборів даних з реальними зображеннями слід віднести Middlebury [9], ETH3D [10].

Окрім використання існуючих реальних наборів даних є можливість генерувати синтетичні дані. Такі датасети можуть бути адаптовані під певні задачі, оточення тощо. Прикладами синтетичних даних є Virtual KITTY [11], SUN3D [12], SceneFlow [13]. Однак навчання лише на синтетичних даних не дозволяє досягти найкращих результатів, і мережі потребують певного доналаштування на даних цільового домену [14].

Одним з альтернативних варіантів є генерування правих зображень, маючи ліві, використовуючи Stereo from Mono алгоритм [14]. Маючи ліве зображення та відповідну істинну карту глибин, праве зображення може бути синтезовано за допомогою обчисленої карти зсувів та певних операцій з пікселями. Проте даний метод призводить до появи пропущених значень у створених зображеннях через оклюзії, які часто не ідеально заповнюються на етапі постобробки синтезованих зображень.

1.2.1 Stereo from Mono Algorithm

Stereo from Mono Algorithm дозволяє з будь-якого набору RGB або RGB-D зображень отримати датасет з лівими I_l, правими I_r зображеннями та картами зсувів D, використовуючи монокулярну мережу для оцінки карти глибин та певні маніпуляції з пікселями зображень. Автори показують, що покращення роботи монокулярної мережі та збільшення кількості тренувальних даних позитивно впливають на результат роботи алгоритму stereo from mono.

Для того, щоб згенерувати праві зображення, спочатку використовується натренерована монокулярна мережа для оцінки карти глибини Z для лівих зображень. Потім автори застосовують передбачені карти відстаней, щоб отримати карти зсувів з використанням різних фокусних відстаней та стереобаз. Різні значеня фокусних відстаней та бейзлайнів обумовлені метою авторів отримати стереонейромережу,

навчену на синтезованих даних, яка буде узагальнюватися на різні датасети з різними параметрами камери. Карти зсувів обчислюються за наступною формулою:

$$D = \frac{sZ_{max}}{Z} \tag{1.1}$$

де s є випадково обраним (рівномірно з діапазону мінімально та максимально можливих значень зсувів $[d_{min},\ d_{max}]$) коефіцієнт масштабування, який гарантує, що створені значення зсувів знаходяться в прийнятному діапазоні.

Далі, використовуючи D, ми можемо синтезувати стереопару за допомогою прямого зміщення [15]. Щоб отримати I_r кожен піксель лівого зображення I_l потрібно перенести на певну кількість пікселів вліво, яка відповідає значенню даного пікселя у карті зсувів D. У такому випадку можуть виникати колізії, коли на один і той самий піксель правого зображення припадає декілька пікселів лівого, а також оклюзії, які відповідають незаповненим значенням у синтезованому правому зображенні. У випадку колізій обирається значення пікселя, яке відповідає більшому значенню зсува так як це означає, що даний піксель розташований ближче до камери, а отже має більшу ймовірність того, що його видно з обох камер, лівої та правої. Для того, щоб усунути оклюзії, які можуть бути присутніми на отриманих зображеннях, автори пропонують виконати перенесення кольору [16] між I_l та випадково обраним зображенням I_b з навчального набору, щоб отримати $\hat{I}_b$ зображення. Потім це зображення використовується для заповнення відсутніх значень у створеному правому зображенні. Короткий опис етапів алгоритму проілюстрований на рис. 1.2. На рис. 1.3 зображені збільшені частини синтезованого правого зображення з рис. 1.3 на яких видно, що пропущені значення в результаті оклюзій заповнені некоректним чином.

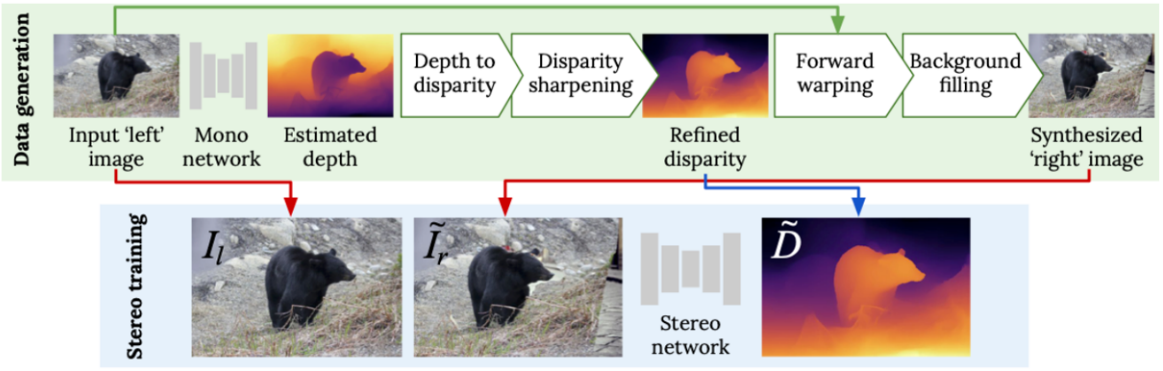


Рисунок 1.2 – Огляд алгоритму Stereo from Mono [14]

Рисунок 1.3 – Приклад заповнення пропущенних значень в результаті оклюзій (червоні квадрати) у синтезованому правому зображенні за допомогою алгоритму Stereo from Mono

1.3 Neural Radiance Fields (NeRF)

NeRF вирішує задачу синтезу нових видів сцени. Маючи певну кількість зображень сцени, відзнятої з різних кутів, NeRF має на меті відновити повну 3D структуру з будь-яких точок зору. Даний метод не є першим, що намагається вирішити поставену задачу. Інші варіанти генерування нових даних включають в себе, наприклад, оптимізацію представлення сцени на основі 3D сітки [17] або методи, що оптимізують нейроменні мережі, які відображать координати xyz у функції відстаней

зі знаком (signed distance functions, SDF) [18, 19]. Однак зазначені методи вимагають істинних 3D даних та є погано масштабованими до отримання зображень великої роздільної здатноті через їх обчислювальну складність. У той же час NeRF має на меті оптимізацію представлення сцени у вигляді неперервної диференційованої функції, що спрощує обчислення.

1.3.1 Neural Radiance Fields (NeRF) - базова модель

NeRF – це представлаення сцени у вигляді певної 5-ти вимірної функції, вхід якої представляється у вигляді 3D координат (x, y, z) та кутів направлення (θ, ϕ), а на виході маємо RGB колір $\vec{c}$ та щільність σ для кожного пікселя [20]. Сама модель нейронної мережі представлена у вигляді багатошарового перцептрона (MLP). Для тренування потрібен набір зображень, який відтворює одну локацію з різних кутів, внутрішні параметри камери та пози камери. Під час тренування мережа синтезує зображення, які порівнюються з оригінальними, мінімізуючи фотометричну помилку між ними.

П'ятивимірне представлення сцени передається в мережу MLP $F_w : (\vec{x}, \vec{d}) \rightarrow (\vec{c}, \sigma)$, де $\vec{x}$ містить 3D координати вибіркових точок, а $\vec{d}$ є тривимірним одиничним вектором, що представляє напрямок променів. Потім вагові коефіцієнти $\vec{w}$ оптимізуються для співставлення кожної вхідної 5D-координати з відповідними об'ємною щільністю та кольором. Щоб отримати узгоджене представлення кількох ракурсів, мережа вираховує щільність σ як функцію лише координат, а вектор кольору $\vec{c}$ як функцію координат та напрямку.

Зображення генерується за допомогою NeRF у такі кроки:

1) Пускаються промені камери через кожний піксель зображення та обираються точки $(a_1,...,a_n)$ уздовж даних променів. Промінь можна обчислити як $r(\vec{t}) = \vec{o} + t\vec{d}$, де де $\vec{o}$ — початок променя, $\vec{d}$ – одиничний вектор напрямку променя. Кожна точка матиме свої унікальні 3D

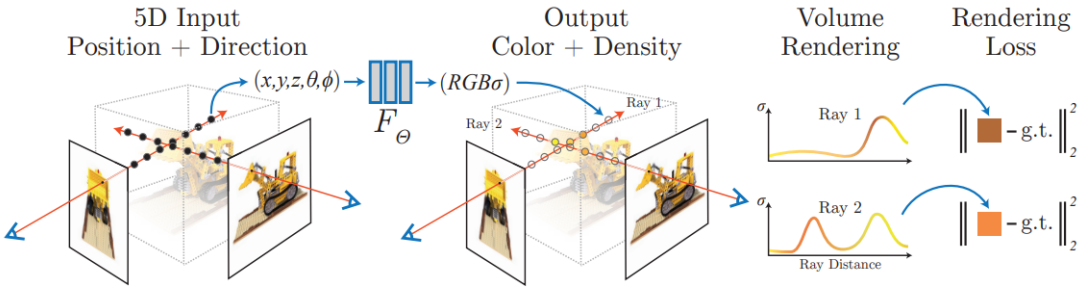


Рисунок 1.4 – Процес синтезу даних за допомогою NeRF [20]

координати з відповідним напрямом променя

2) Отримані точки передаються до мережі, з якої отримуються відповідні колір та щільність для кожної з точок

3) Далі застосовуються класичні техніки рендерингу [21], щоб отримати кінцеві значення кольорів пікселів, через які проходять промені. Формула для обчислення кольору виглядає наступним чином:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt \tag{1.2}$$

де t_n, t_f – мінімальне та максимальне значення, які може приймати t, $r(t)$ – вхідний промінь, $\sigma(r(t))$ – об'ємна щільність, яку також можна інтерпретувати як ймовірність закінчення променя в точці t, $c(r(t), d)$ – колір променя в точці t.

Змінну $T(t)$ у вищенаведеному рівнянні можна обчислити, використовуючи формулу:

$$T(t) = exp(-\int_{t_n}^{t} \sigma(r(s)ds) \tag{1.3}$$

$T(t)$ - це певний коефіцієнт, який характеризує проникність променю до точки t тобто ймовірність того, що промінь пройде від точки t_n до t, не стикаючись з іншими частинками.

Процес рендерингу проілюстрований на рисунку 1.4

NeRF має ряд недоліків та переваг. До недоліків слід віднести час тренування (1 - 2 дні) та той факт, що дана модель не узагальнюється на інші сцени тобто її доречнно використовувати для генерації нових

зображень саме тієї локації, на зображеннях якої вона тренувалася. Також, зображення, отримані за допомогою стандартної моделі NeRF у випадку сцен, розміром з кімнату або більше, є не чіткими та мають певні артефакти, наприклад, шум.

Наразі існує багато різних підходів Neural Radiance Fields, які застосовується у різних областях. Серед них можна виділити декілька груп, наприклад, моделі NeRF, які можна застосовувати для редагування сцени [22, 23], сегментації зображень [24, 25], оцінки карт глибин, 3D реконструкція сцени [26, 27, 28]. Існують види даної моделі, які не потребують на вхід пози камери, а знаходять і оптимізують її в процесі тренування [29, 30],

1.3.2 Bundle-Adjusting Neural Radiance Fields (BARF)

Дана модель є підвидом NeRF та розширює можливості оригінальної моделі. Одним з головних обмежень NeRF є те, що вона потребує ідеальні пози камери. Зазвичай ми маємо неідеальні пози камери або взагалі не маємо їх у своєму розпорядженні. Для того, щоб отримати додатково зовнішні параметри зображень потрібно скористатися додатковими техніками, наприклад, COLMAP [31]. Проте методи, за допомогою яких можна оцінити пози камери сенсорів, є неточними [30]. BARF дозволяє обійти дане обмеження, вводячи додаткові параметри, які описують зміну положення камери, до функції, яка оптимізується. Також ще однією особливістю даної моделі є зміна застосування позитційного кодування.

Означення 1.1. Позиційне кодування – це детерміноване відображення вхідних 3D-координат $\vec{x}$ на вищі степені різних синусоїдальних баз частот.

У випадку синтезування нових видів за допомогою NeRF, дана техніка відіграє важливу роль. Адже саме за допомогою її використання

вдається досягти чітких та розбірливих згенерованих картинок [20]. Позиційне кодування дозволяє нейронним мережам, які представляють сцени у вигляді координат, подавати сигнали вищої частоти та досягати швидшої збіжності [32]. Автори провели певні дослідження щодо застосування повного позиційного кодуввання або не використання його взагалі, та прийшли до висновку, що кращим варіантом є поступове застосування даної функції, зважаючи на епоху тренування. Вони показали, що застосування повного позиційного кодування призводить до шумного сигналу і можливих локальних мінімумів.

1.3.3 Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo (NerfingMVS) [33]

Метод грунтується на базовій моделі NeRF та використовує керовану оптимізацію. Ідея полягає в тому, як ефективно застосувати додаткову інформацію, отриману на перших кроках оптимізації NeRF. Модель спочатку оптимізує монокулярну мережу для оцінки карт глибин на цільовій сцені шляхом донавчання її на розріджених картах відстаней, отриманих за допомогою structure from motion (SfM) з COLMAP [31]. Процедура обрання точок на променях при оптимізації NeRF базується на отриманій карті помилок синтезованих зображень. Таким чином карти глибин уточнюються і отриманий результат є більш точним та чітким, порівняно з результатами, отриманими за допомогою оригінальної моделі NeRF. Огляд мережі зображений на рисунку 1.5. Хоча основна ціль даної моделі - це покращення отриманих карт глибин, автори NerfingMVS показують, що такий спосіб не погіршує спроможність NeRF генерувати нові зображення, а покращує її.

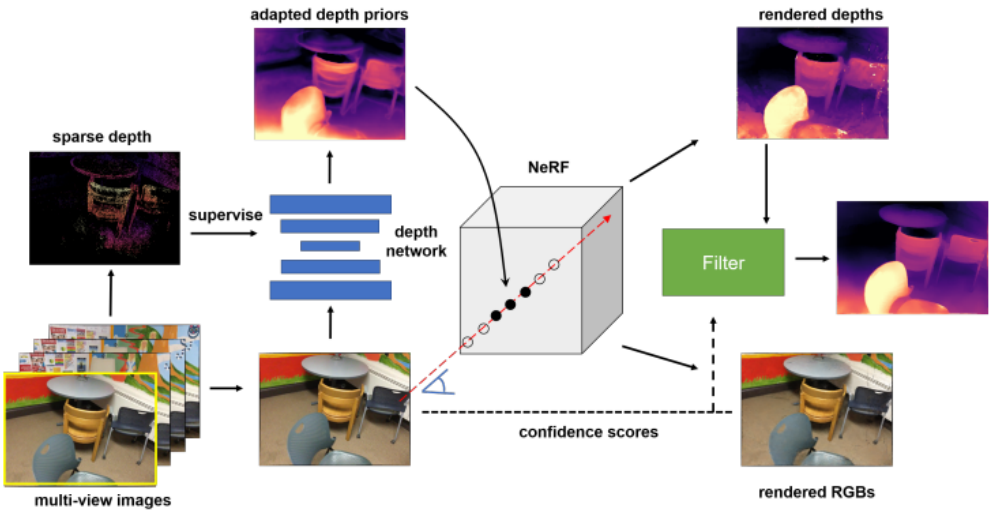


Рисунок 1.5 – огляд моделі NerfingMVS [33]

1.3.4 Dense Depth Priors for Neural Radiance Fields from Sparse Input Views (DDPNeRF)

Однією з проблем Neural Radiance Fields є велика кількість зображень, які потрібно подавати на вхід мережі. Це є обмеженням адже в нашому розпорядженні може бути всього декілька картинок сцени, і ми хочемо також отримувати чіткий результат. Автори роботи [34] вирішують дану проблему, використовуючи COLMAP [31] та генеруючи щільні карти глибин разом з картами невизначеності, які далі використовуються для оптимізації NeRF.

На вхід мережі подаються спеціально підготовлені картинки сцени та розріджені карти глибин, які отримані за допомогою SfM з COLMAP. Далі автори використовують модель, яка доповнює розріджені карти глибин та знаходять карти невизначеності, які далі використовуються для оптимізації NeRF. Таким чином даний метод дозволяє генерувати нові зображення сцени, розміром з кімнату, використовуючи як вхід всього 18-36 зображень. DDPNeRF є схожим на NerfingMVS [33] проте використовує заповнення карт відстаней, отриманих з COLMAP, а також обчислює та використовує карти невизначеності. Огляд моделі зображений на рисунку 1.6

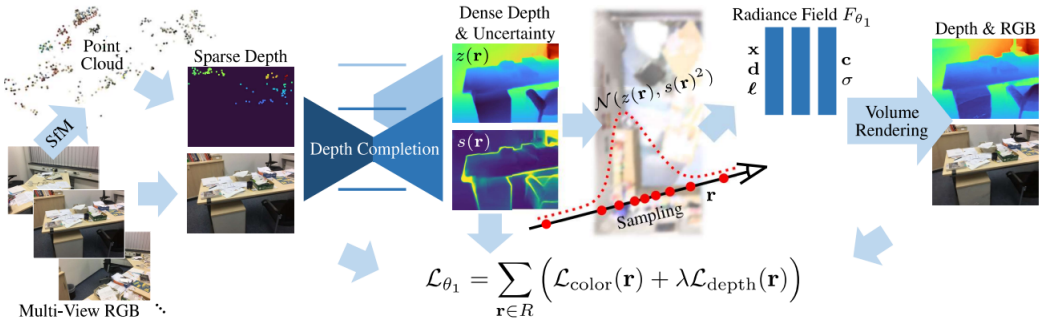


Рисунок 1.6 – огляд моделі DDPNeRF [34]

1.3.5 Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo (MVSNeRF)

Як було зазначено на початку, одним з недоліків NeRF є їх неузагальненість на інші сцени. Існують, наприклад, датасети, які складаються з багатьох різних сцен, тож для таких випадків даний недолік є суттєвим обмеженням. Модель MVSNeRF [35] частково вирішує дану проблему. Використовуючи дану модифікацію NeRF, є можливість не тренувати модель для кожного датасету окремо з самого початку, а натренувати на одному наборі даних, а потім дотреновувати на іншому. Автори зазначають, що достатньо навіть 15 хв. часу, щоб отримати порівняно гарні результати для інших сцен. Автори роботи тренували модель на DTU [36] датасеті та застосовували її для донавчання на інших датасетах таких як, наприклад, LLFF [ref] чи Realistic Synthetic [ref], щоб перевірити здатність отриманої моделі до узагальнення.

Ідея даної роботи полягає у застосуванні cost volume для кращої оцінки геометрії сцен, а також модель використовує 3D CNNs. Cost volume відповідає відстані між обчисленими фічами вхідних зображень для різних гіпотез щодо передбаченої глибини базовового зображення. Зокрема, використовуючи 3D CNN, автори реконструюють (виходячи з cost volume) певне нейронне представлення сцени, яке складається з нейронних характеристик кожного вокселя, які кодують інформацію про

геометрію та зовнішній вигляд локальної сцени.

Спочатку мережа будує cost volume, використовуючи перетворення 2D ознак зображень, далі застосовуються 3D конволюційні шари для реконструкції об'єму з нейронними характеристиками для кожного вокселя. Для отримання щільності σ та RGB кольору використовується, як і в оригінальній моделі NeRF, багатошаровий перцептрон. Огляд усієї моделі зображений на рисунку 1.7.

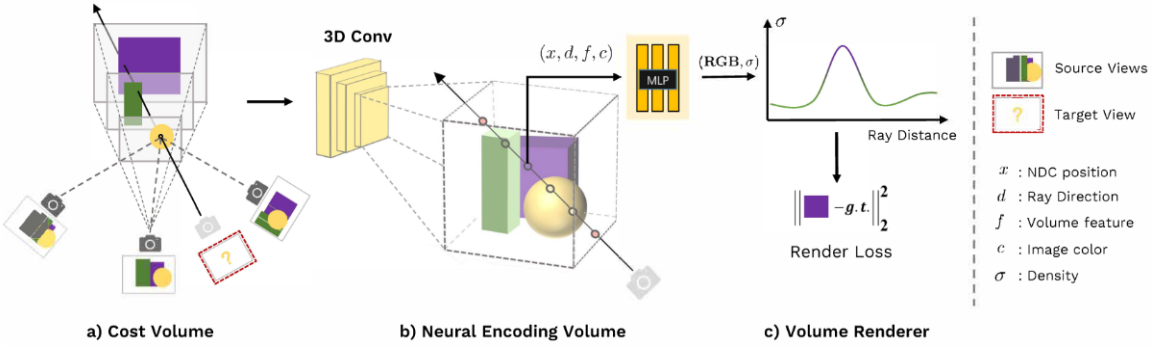


Рисунок 1.7 – огляд моделі MVSNeRF [35]

Отже, порівняно з іншими роботами, пов'язаними з синтезом зображень за допомогою NeRF, дана модель використовує переваги MVS архітектури, що дає змогу краще знаходити різні відповідності між зображеннями. А також поєднання MVS + 3D CNN дають змогу узагальнювати модель на інші датасети.

1.3.6 Implicit Mapping and Positioning in Real-Time (iMAP)

Усі моделі Neural Radiance Fields, які були перелічені, здатні працювати тільки офлайн. iMAP - це перший метод, побудований на основі NeRF, який може служити єдиним представленням сцени у системі SLAM, яка використовується у реальному часі для портативної RGB-D камери [37]. Мережа навчена працювати в реальному часі без попередніх даних, будуючи щільну, специфічну для конкретної сцени неявну

3D-модель щільності та кольору, яка також використовується для трекінгу камери.

Автори iMAP пропонують відображати сцену, поступово оптимізуючи ваги мережі та пози камери на основі декількох спостережень та застосовуючи певні методи вибору ключових фреймів для оптимізації. Опис системи зображений на рисунку 1.8.

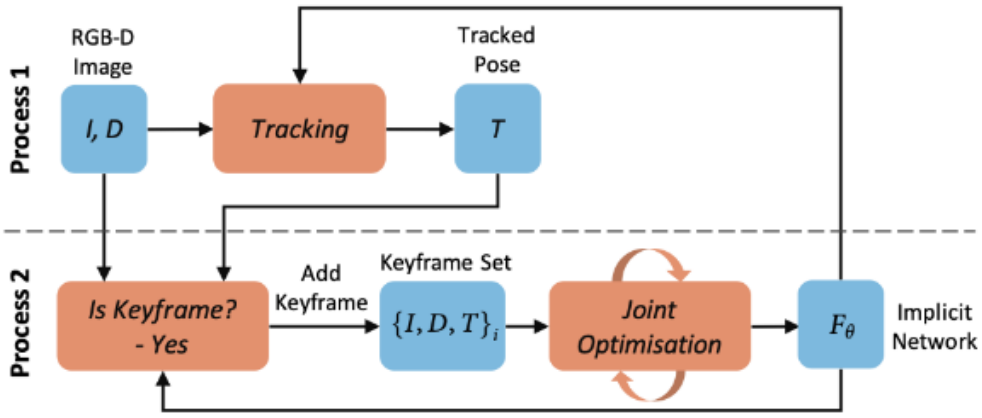


Рисунок 1.8 – огляд моделі iMAP [37]

iMAP складається з двох блоків, які працюють одночасно: перший блок, трекінг, відповідвє за оптимізацію пози камери поточного кадру відносно заблокованої мережі; другий блок, відображення, одночасно оптимізує ваги мережі та пози камери обраних кадрів. Фрейми обираються, використовуючи інформаційний виграш (information gain). Дані картинки є певним банком пам'яті, щоб уникнути забуття мережі.

1.3.7 Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes (NSFF)

Усі вищезгадані підвиди NeRF мають одне спільне припущення – локація є статичною. Автори моделі NSFF [38] видозмінили базове представлення сцени як NeRF, враховуючи динамічність локацій. Нове представлення сцени, Neural Scene Flow Fields, моделює динамічну сцену

як змінну в часі безперервну функцію вигляду оточення, геометрії та руху 3D-сцени. Такий підхід дозволяє інтерполювати зміни як у просторі, так і у часі. Ще однією відмінністю від інших моделей NeRF є те, що, крім кольору та щільності, NSFF передбачає 3D динаміку сцени. Для обробки динамічнтх дизоклюзій, додатковим вихідним параметром NSFF є також ваги дизоклюзії. Отже, за допомогою даної моделі можна, наприклад, синтезувати нові картинки локації з динамічним об'єктом, залишаючи рух динамічного об'єкта, а сама сцена при цьому буде нерухомою.

1.3.8 LENS: Localization enhanced by NeRF synthesis

LENS – це підхід до вирішення задачі локалізації, використовуючи синтез зображень за допомогою Neural Radiance Fields [39]. Автори показують, що додаткове генерування даних покращує точність регресії пози камери. Щоб уникнути створення нових видів у невідповідних місцях, в даному методі розташування віртуальних камер обирається із внутрішнього 3D представлення NeRF. Такий підхід дозволяє майже нескінченно генерувати навчальні дані. На момент публікації статті підхід удосконалив наявні методи із зниженням помилки на 60% для наборів даних Cambridge Landmarks [40] та 7 scenes [41]. Отже, результуюча точність стає порівнянною зі структурними методами без будь-яких модифікацій архітектури чи обмежень адаптації домену [39]. Структурні методи зазвичай дають змогу досягти найкращих результатів для задач локалізації, проте вони потребують великих обчислювальних можливостей та затрат пам'яті, що унеможливлює їх застосування у режимі реального часу.

У кінці всіх експериментів, автори прийшли до висновку, що точність регресії пози камери здебільшого обмежена відносно невеликими та упередженими наборами даних, а не здатністю моделі регресії пози вирішити завдання локалізації. Для синтезу нових даних була

використана модель NeRF-W [42].

Модель LENS представлени на рис. 1.9.

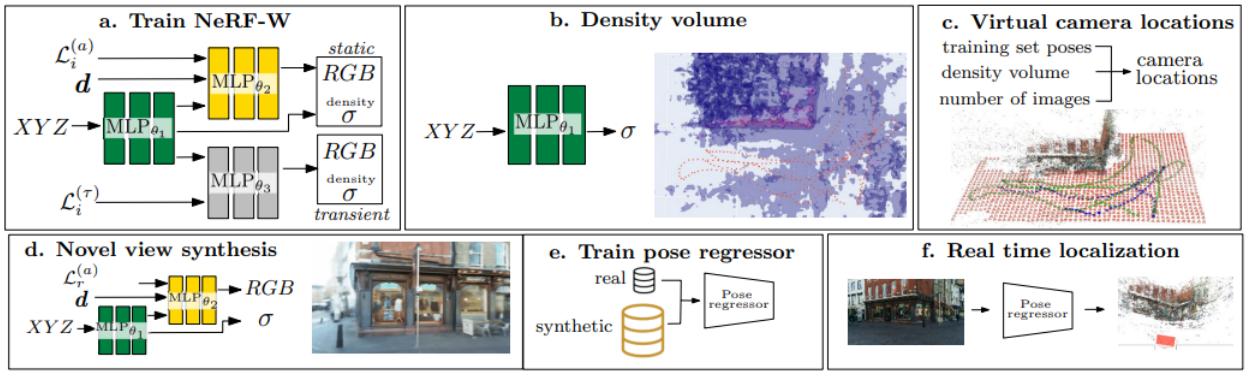


Рисунок 1.9 – Огляд моделі LENS [39]

Спочатку NeRF-W навчається на доступних реальних зображеннях (a.). Далі навчена модель використовується для виявлення точок, які мають високу щільність в сцені (b.). Пози реальних зображень і місця високої щільності використовуються для створення віртуальних розташувань камер (c.), для яких виконується синтез нових видів (d.). Далі реальні та синтетичні набори даних разом використовуються для навчання регресора пози (e.).

Отже, у даному методі розроблена технологія використання NeRF для нескінченного синтезу даних під час тренування іншої мережі. Тобто це можна вважати додатковою аугментацією вхідних даних. Саме дана модифікація у ланцюжку роботи методу дає можливість значно покращити точність вихідної моделі (до 60% відносно інших моделей).

Висновки до розділу 1

Проблема кількості наявних датасетів, які підходять під усі характеристики для тренування стереонейромереж, є важливою. Однією

з альтернатив є використання синтезованих датасетів, проте тільки їх застосування в значній мірі не дає покращення якості моделі. А існуючий метод, який дозволяє з будь-якого набору лівих зображень генерувати праві, має ряд недоліків, пов'язаних з появою пропущених значень в результаті оклюзій у згенерованих зображеннях, які неправильно заповнюються (див. рис. 1.3). На сьогоднішній день є багато різних моделей Neural Radiance Fields, які дозволяють синтезувати фотореалістичні нові представлення сцени, використовуючи для тренування, в оригінальному варіанті, тільки набір зображень та пози камери. Даний підхід потенційно може бути використаний для генерування потрібних тренувальних даних, проте раніше в літературі такий спосіб не було досліджено. У даній роботі пропонується можливість генерування стереоданих, використовуючи NeRF.

2 ГЕНЕРУВАННЯ ЗОБРАЖЕНЬ ЗА ДОПОМОГОЮ NERF

2.1 Постановка задачі

Тренування стереонейромереж для оцінки карт глибин може відбуватися декількома способами. Перший з них – це навчання з учителем, коли ми маємо достатню кількість даних з істинними значеннями відстаней до об'єктів чи картами зсувів для тренування. Іншим варіантом є навчання без учителя, який використовується за відсутності істинних навчальних даних. Останній спосіб є досить розповсюдженим адже важко зібрати велику кількість даних з повними та коректними картами глибин. Методи навчання без учителя використовують переваги проективної геометрії. В процесі даного навчання оптимізуються різні види фотометричних функцій втрат, наприклад, функції, які включають в себе обрахунок SSIM (structure similarity index) [43, 44, 45], а також функції, які обраховують помилку між певними спроектованими значеннями глибин та зображеннями на інші тренувальні дані та власне картами глибин та зображеннями на які відбувалася проекція [43, 45]. Також можлива ситуація об'єднання даних двох варіантів навчання нейромереж в один. Такий варіант можливий, наприклад, при виникненні потреби в доналаштування моделі на даних з цільової області за відсутності істинних карт глибин.

Розгляньмо ланцюжок тренування, використовуючи навчання без учителя. У такому варіанті ми не знаємо істинні дані, і кінцевий результат зазвичай є гіршим за результати тренування з учителем. Отже, виникає питання чи можна згенерувати потрібні тренувальні дані та навчати стереонейромережу у режимі навчання з учителем? У даній роботі розглянуто можливість синтезування тренувальних даних, використовуючи технологію NeRF, та їх використання у тренуванні стереонейромережі. Також за допомогою NeRF є можливість згенерувати

будь-яку кількість навчальних даних, що може призвести до покращення результату.

Отже, стадартний ланцюжок навчання стереонейромереж для оцінки карт глибин виглядає як показано на рис. 2.1. У роботі розглядається варіант тренування, який також включає в себе додаткову підготовку даних. Пропонована схема тренування зображена на рис. 2.2.

Рисунок 2.1 – Стандартний ланцюжок навчання стереонейромереж для оцінки карт глибин (тренування з учителем / тренування без учителя)

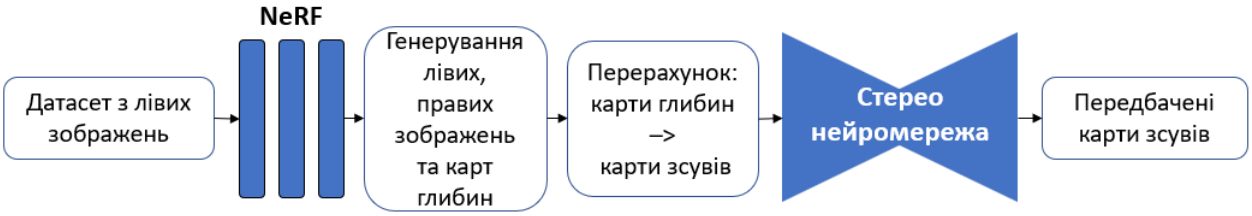


Рисунок 2.2 – Пропонований ланцюжок навчання стереонейромереж для оцінки карт глибин

Для вирішення поставленої задачі було зроблено такі кроки:

1) Обрана модель NeRF для дослідження – BARF.

2) Були зроблені певні модифікації моделі, а саме: введена додаткова функція втрат для оптимізації отриманих карт глибин та введена нова вхідна змінна – час t.

3) Для повної оцінки отриманих результатів було проведено декілька порівняльних тренувань стереонейромережі:

а) Тренування з учителем, використовуючи істинні дані ScanNet

б) Тренування з учителем, використовуючи згенеровані навчальні дані за допомогою NeRF

в) Тренування без учителя, використовуючи істинні зображення ScanNet

2.2 Оцінка карт глибин зі стереозображень

Задача стереозору, як зазначалося у першому розділі роботи, полягає у відшуканні 3D структури об'єкту чи сцени за наявними стерео 2D зображеннями. Перевагою оцінки карт глибин з набору стереозображень над монокулярним передбаченням карт відстаней є те, що використання інформації з двох зображень, лівого та правого, дозволяє будувати, наприклад, певні геометричні обмеження, які призводять до покращення результатів. Одним з недоліків монокулярного зору є некоректна оцінка масштабу або, наприклад, виникнення оклюзій через велику кількість об'єктів у сцені [46]. Використання стереозображень у певній мірі дозволяють вирішити наведені обмеження монокулярної оцінки глибини. Проте у стереосистемі також є недоліки, наприклад, необхідність у наявності двох камер та їх калібрування. Стереонейромережі мають на меті відшукання відповідності між лівими та правими зображеннями сцени, передбачаючи карти зсувів. Значення зсуву відповідає різниці в місці розташування зображення об'єкта, який бачать ліва та права камери. Приклад стереозображень та відповідної їм карти зсувів зображений на рис 2.3. Далі за допомогою відомих параметрів камери карти зсувів можуть бути перераховані у відповідні їм карти глибин.

У загальному випадку стереонейромережі вирішують задачу регресії. На вхід моделі подаються 2 зображення, які відповідають лівій та правій камерам. Такі зображення є ректифікованими та розташовані на певній відстані одне від одного. Ректифіковані зображення – це зображення, які розташовані так, що права камера зміщена відносно лівої

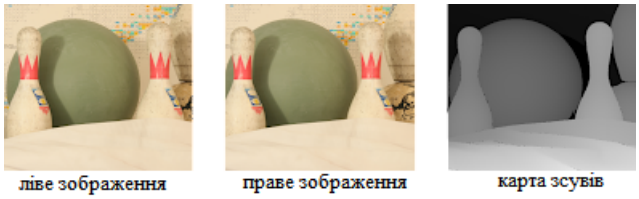


Рисунок 2.3 – Приклад стереозображень та відповідної їм карти зсувів

тільки по горизонталі, а усі епіполярні лінії є горизонтальними. У випадку тренування з учителем, додатково подаються істинні дані – карти зсувів. На виході мережа передбачає карти зсувів, які, за допомогою відомих параметрів камер, перераховуються у відповідні карти глибин. Розповсюдженим варіантом архітектури стереонейромереж є архітектури виду U-Net [47], тобто мережі вигляду encoder-decoder, які використовують 2D або 3D конволюційні шари (CNN). Прикладом такої моделі може слугувати DispNet [13], архітектура якої зображена на рис. 2.4. Як видно зі схеми дана мережа складається з поєднаних 2D конволюційних шарів з шарами MaxPooling2D, за ланцюжком яких слідує частина декодера, яка відображена у шарах Deconvolution2D та Upsampling2D. Також присутні зв'язки skip connection, які передають інформацію з минулих шарів нейромережі.

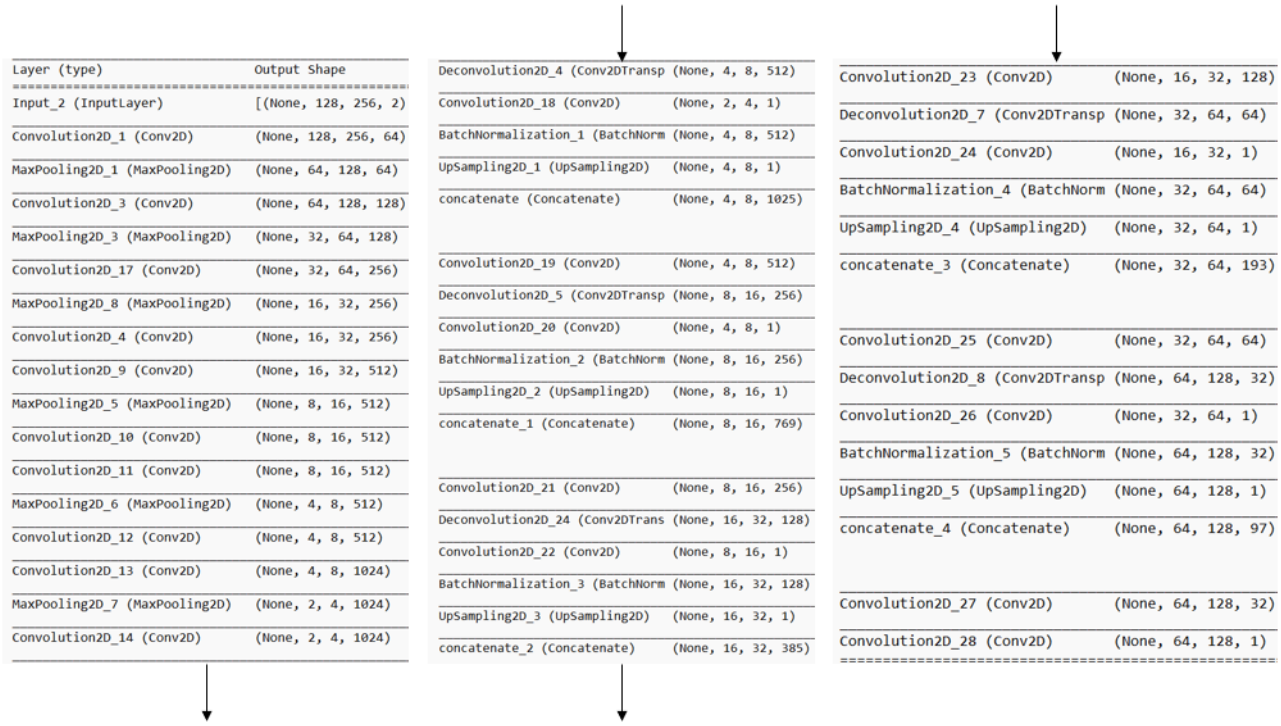
Функції втрат відповідно поділяються на дві категорії: функції, які застосовуються при навчанні з учителем та функції, які використовуються у випадку тренування без учителя.

1) Тренування з учителем

Найпростіший випадок функції втрат для оптимізації отриманих карт зсувів – є розрахунок помилок між істинними та передбаченими картами значеннями як функцію відстані:

$$L = \frac{1}{N} \sum_{i=1,\vec{N}} D(d_i, \hat{d}_i) \tag{2.1}$$

де N – кількість тренувальних даних, D – функція відстані (наприклад, L_1, L_2 [48, 49]), $\hat{d}$ – істинні карти зсувів, d - передбачені карти зсувів



Layer (type)	Output Shape
Input_2 (InputLayer)	[(None, 128, 256, 2)
Convolution2D_1 (Conv2D)	(None, 128, 256, 64)
MaxPooling2D_1 (MaxPooling2D)	(None, 64, 128, 64)
Convolution2D_3 (Conv2D)	(None, 64, 128, 128)
MaxPooling2D_3 (MaxPooling2D)	(None, 32, 64, 128)
Convolution2D_17 (Conv2D)	(None, 32, 64, 256)
MaxPooling2D_8 (MaxPooling2D)	(None, 16, 32, 256)
Convolution2D_4 (Conv2D)	(None, 16, 32, 256)
Convolution2D_9 (Conv2D)	(None, 16, 32, 512)
MaxPooling2D_5 (MaxPooling2D)	(None, 8, 16, 512)
Convolution2D_10 (Conv2D)	(None, 8, 16, 512)
Convolution2D_11 (Conv2D)	(None, 8, 16, 512)
MaxPooling2D_6 (MaxPooling2D)	(None, 4, 8, 512)
Convolution2D_12 (Conv2D)	(None, 4, 8, 512)
Convolution2D_13 (Conv2D)	(None, 4, 8, 1024)
MaxPooling2D_7 (MaxPooling2D)	(None, 2, 4, 1024)
Convolution2D_14 (Conv2D)	(None, 2, 4, 1024)

Deconvolution2D_4 (Conv2DTransp	(None, 4, 8, 512)
Convolution2D_18 (Conv2D)	(None, 2, 4, 1)
BatchNormalization_1 (BatchNorm	(None, 4, 8, 512)
UpSampling2D_1 (UpSampling2D)	(None, 4, 8, 1)
concatenate (Concatenate)	(None, 4, 8, 1025)
Convolution2D_19 (Conv2D)	(None, 4, 8, 512)
Deconvolution2D_5 (Conv2DTransp	(None, 8, 16, 256)
Convolution2D_20 (Conv2D)	(None, 4, 8, 1)
BatchNormalization_2 (BatchNorm	(None, 8, 16, 256)
UpSampling2D_2 (UpSampling2D)	(None, 8, 16, 1)
concatenate_1 (Concatenate)	(None, 8, 16, 769)
Convolution2D_21 (Conv2D)	(None, 8, 16, 256)
Deconvolution2D_24 (Conv2DTrans	(None, 16, 32, 128)
Convolution2D_22 (Conv2D)	(None, 8, 16, 1)
BatchNormalization_3 (BatchNorm	(None, 16, 32, 128)
UpSampling2D_3 (UpSampling2D)	(None, 16, 32, 1)
concatenate_2 (Concatenate)	(None, 16, 32, 385)

Convolution2D_23 (Conv2D)	(None, 16, 32, 128)
Deconvolution2D_7 (Conv2DTransp	(None, 32, 64, 64)
Convolution2D_24 (Conv2D)	(None, 16, 32, 1)
BatchNormalization_4 (BatchNorm	(None, 32, 64, 64)
UpSampling2D_4 (UpSampling2D)	(None, 32, 64, 1)
concatenate_3 (Concatenate)	(None, 32, 64, 193)
Convolution2D_25 (Conv2D)	(None, 32, 64, 64)
Deconvolution2D_8 (Conv2DTransp	(None, 64, 128, 32)
Convolution2D_26 (Conv2D)	(None, 32, 64, 1)
BatchNormalization_5 (BatchNorm	(None, 64, 128, 32)
UpSampling2D_5 (UpSampling2D)	(None, 64, 128, 1)
concatenate_4 (Concatenate)	(None, 64, 128, 97)
Convolution2D_27 (Conv2D)	(None, 64, 128, 32)
Convolution2D_28 (Conv2D)	(None, 64, 128, 1)

Рисунок 2.4 – Архітектура DispNet

Також до функції втрат може додаватися функція регуляризації або гладкості, яка накладає додаткові локальні або глобальні обмеження [50]

2) Тренування без учителя

Функції втрат у випадку навчання без учителя грунтуються на використанні функцій подібності та проективної геометрії. У загальному випадку функція втрат виглядає наступним чином [50]:

$$L = \frac{1}{N} \sum F(I_{ref}, I_{warp}) \tag{2.2}$$

де N – кількість тренувальних даних, F – функція подібності, I_{ref} – референсне зображення, на яке робиться проекція, I_{warp} – спроектоване зображення на референсне

Проекція обраховується, використовуючи матрицю внутрішніх параметрів камери K, матрицю повороту R та вектор зміщення T між двома кадрами. Формула для обчислення виглядає наступним чином [45]:

$$z^{'}p^{'} = KRK^{-1}zp + KT \tag{2.3}$$

де p, z - однорідні піксельні координати та глибина, а $p^{'}$, $z^{'}$ їх відповідні спроектовані значення

Також додатковими "ключами"до оптимізації карт відстаней можуть бути функції втрат, які регулюють гладкість отриманих карт зсувів [50, 45].

Перевагою end-to-end навчання та використання стереонейромереж виду енкодер-декодер є їх простота в оптимізації та швидкодія. Проте для застосування таких видів нейромереж потрібна велика кількість навчальних даних. Це приблизно десятки тисяч зображень. Такий обсяг даних є достатньо великим, що є суттєвим недоліком такого типу нейромереж, адже зібрати або знайти у відкритому доступі зазначену кількість картинок з відповідними картами відстаней та потрібними параметрами камери є складною задачею.

2.3 Bundle-Adjusting Neural Radiance Fields

BARF - це один з видозмінених варіантів базової моделі NeRF, який, окрім оптимізації синтезованих зображень, у процесі тренування також оптимізує пози камери. Як було зазначено у розділі 1, однією з головних проблем NeRF є те, що на вхід моделі очікуються ідеальні пози камери, які не завжди є доступними.

Маючи вхідні зображення $(I_1,...,I_M)$, мета даної моделі - оптимізувати NeRF та пози камери $(p_1,...,p_M)$, мінімізуючи наступну функцію втрат:

$$L = \sum_{i=1}^{M} \sum_{u} ||\hat{I}_i(u; p_i, w) - I_i(u)||_2^2 \tag{2.4}$$

де w - параметри мережі, що також залежать від напрямку променів, u - координати пікселів, $\hat{I}_i(u; p_i, w)$ - синтезований RGB колір в пікселі u, який відповідає i-му зображенню

Ще одна відмінність BARF від початкової моделі NeRF [20] полягає у зміні застосування позиційного кодування. Позначимо позиційне кодування як функцію:

$$\gamma : \mathbb{R}^3 \to \mathbb{R}^{3+6L} \tag{2.5}$$

де L – частотний базис

Дана функція визначається наступним чином:

$$\gamma(\vec{x}) = [\vec{x}, \gamma_0(\vec{x}), ..., \gamma_{L-1}(\vec{x})] \in \mathbb{R}^{3+6L} \tag{2.6}$$

де k-та компонента записується у вигляді $\gamma_k(x) = [cos(2^k \pi \vec{x}), sin(2^k \pi \vec{x})] \in \mathbb{R}^6$.

Ідея BARF полягає у застосуванні зваженого позиційного кодування тобто k-та компонента у даній моделі рахується як:

$$\gamma_k(x, \alpha) = w_k(\alpha)[cos(2^k \pi \vec{x}), sin(2^k \pi \vec{x})] \tag{2.7}$$

де ваги $w_k(\alpha)$ визначаються наступним чином:

$$w_k(\alpha) = \begin{cases} 0, & \text{if } \alpha < \text{k} \\ \frac{1-cos((\alpha-k)\pi)}{2}, & \text{if } 0 \leqslant \alpha \text{ - k} < 1 \\ 1, & \text{if } \alpha - k \geqslant 1 \end{cases} \tag{2.8}$$

Параметр $\alpha \in [0, L]$ у формулі 2.4 є контрольованим та пропорційним процесу оптимізації тобто його величина залежить від етапу тренування мережі. Починаючи з перших епох тренування, поступово активізується кодування вищих діапазонів до моменту $\alpha = L$, що еквівалентно повному позиційному кодуванню. Це дозволяє BARF знаходити більш оптимальні рішення для передбачення поз камери [30]. Приклад з порівнянням синтезованих зображень для моделей з різним позиційним кодуванням проілюстрований на рис. 2.5.

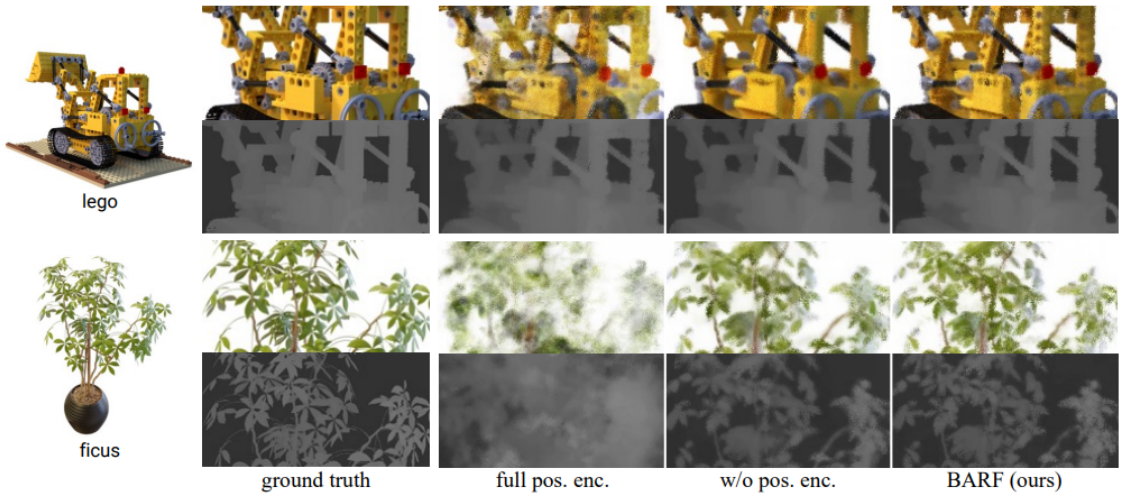


Рисунок 2.5 – Приклад результатів генерування зображень за допомогою BARF з різним типом позиційного кодування [30]. Верхні рядки відповідають синтезованим зображенням, а нижні - передбаченим картам глибин

2.3.1 Функція втрат для оптимізації оцінки карт глибин

Оригінальна модель NeRF та усі її покращення в якості одного з виходів мережі мають щільність σ, яку можна інтерпретувати як величину непрозорості об'єкту. Використовуючи отриману щільність, можна вирахувати відстані до об'єктів. У процесі навчання NeRF передбачені карти глибин додатково не оптимізуються. Це призводить до того, що результуючі карти відстаней зазвичай мають певні недоліки, наприклад, можуть з'являтися літаючі об'єкти як зображено на рис. 2.6. Також передбачені відстані є такими, що мають достатні розбіжності з істинними даними (помилки досягають порядку 10-20%, як буде розглянуто у 3 розділі роботи). Для покращення точності передбачених карт глибин введемо додаткову функцію втрат, яка визначається як MSE помилка між істинними значеннями відстаней до об'єктів та передбаченими. Для її використання дані повинні містити істинні карти глибин. Дана функція втрат не є новою та часто використовується у

монокулярних та стереонейромережах, мета яких – це оцінка карт глибин. Також замість MSE використовують, наприклад, L1 відстань.

Рисунок 2.6 – Приклад передбаченої карти зсувів моделлю BARF. Червоні кола відповідають "літаючим об'єктам"

У статті [50] проілюстровано використання зазначених функцій втрат. Таким чином, отримуємо наступну функцію, яку потрібно оптимізувати:

$$L_D = ||D_{gt} - \hat{D}||_2^2 \tag{2.9}$$

де D_{gt} – істинні карти глибин, $\hat{D}$ – карти глибин, отримані за допомогою BARF

Отже, результуюча функція втрат виглядає наступним чином:

$$L = L_F + L_D = ||I(u; p, w) - I(u)||_2^2 + ||D_{gt} - \hat{D}||_2^2 \tag{2.10}$$

де u - координати пікселів, $\hat{I}(u; p_i, w)$ - синтезований RGB колір в пікселі u, D_{gt} – істинні карти глибин, $\hat{D}$ – карти глибин, отримані за допомогою BARF

2.3.2 Зміна вхідних параметрів моделі

Однією з особливостей реальних даних є зміна освітленості. Зазвичай наявні дані мають статичне освітлення, яке спостерігається у всьому датасеті. Зміна освітленості сцени може виникати з різних причин, наприклад, від зовнішнього джерела такого як сонце або від фар автомобілів. Також дані зміни можуть залежати від локальних змін у локаціях наборів даних, наприклад, включення/виключення джерела освітлення у приміщеннях одного датасету. Зміна освітленості сцен може негативно вплинути на результати роботи мережі для оцінки карт глибин [51]. Особливо це стосується нейромереж для оцінки карт відстаней, які навчаються у режимі без учителя так як у такому випадку застосовуються проектування між зображеннями та функції втрат, наприклад, SSIM або L_1 відстань між референсними зображеннями та спроектованими на них. Зазначені функції втрат залежать від інтенсивності зображень. У випадку навчання без учителя працює припущення про однакову яскравість даних [51], але якщо у наборі даних освітлення змінюється, то це призводить до порушення даного припущення. У роботі [51] автори наводять приклад роботи монокулярної мережі MonoDepth2 [52] для оцінки карт глибин, що навчалася у режимі без учителя. Приклад результатів наведений на рис. 2.7. З отриманих карт відстаней видно, що погане або змінне освітлення призводить до появи великих областей з пропущенними значеннями та до негладкості.

За допомогою NeRF можна спробувати додатково моделювати освітлення, що дало б змогу зробити стереонейромережі більш стійкими до даних змін. Для того, щоб мати змогу моделювати зміну освітлення певної сцени, додаймо до вхідних параметрів моделі змінну часу t. Даний параметр буде відповідати номеру картинки, що подається для тренування NeRF. Індекс зображення має лінійний зв'язок з часом тому що набори даних, з якими ми маємо справу, це відео, відзняті з

Рисунок 2.7 – Приклад передбачених карт глибин, використовуючи MonoDepth2 [52] у випадку поганого або змінного освітлення [51]. Синім штрихованим прямокутником виділено темну область зображення. Червоним штрихованим прямокутником виділено область зі змінним освітленням, яке відображено у збільшених червоних прямокутниках, де t - один кадр, $t+1$ - наступний

фіксованою частотою кадрів. Також всі індекси $t_i, i = (1,\vec{M})$, де M – розмір датасету, попередньо нормалізуються до інтервалу [0,1] та до них застосовується позиційне кодування, як і у випадку для вхідних 3D координат та векторів напавлення вхідних променів. Отже, змінений вигляд моделі BARF виглядає наступним чином:

$$F_w : (\vec{x}, \vec{d}, \vec{t}) \rightarrow (\vec{c}, \sigma), \tag{2.11}$$

де $\vec{x}$ містить 3D координати вибіркових точок, $\vec{d}$ є тривимірним одиничним вектором, що представляє напрямок променів, $\vec{t} \in [0,1]$ містить нормовані індекси усіх картинок датасету.

У статичній сцені об'єкти з плином часу не змінюють своє положення, проте умови освітленості сцени можуть змінюватися. Отже, зміна освітленості відповідає тільки зміні значень інтенсивності та не змінює щільність. Це означає, що для моделювання освітленості сцени,

щільність σ буде залежити тільки від 3D координат вхідних точок, а RGB колір $\vec{c}$ залежитиме від 3D координат, вектору направлення $\vec{d}$, а також часу $\vec{t}$. Проте у даній роботі розглянуто вплив ще одного варіанту додавання змінної часу до вхідних параметрів. Зазначена модифікація передбачає, що t впливає і на передбачений NeRF колір, і на щільність.

Висновки до розділу 2

У даному розділі розглянуто стандартні варіанти тренування глибоких нейромереж для оцінки карт глибин зі стереозображень та запропоновано модифікований ланцюжок підготовки даних та тренування стереонейромереж, який передбачає попередню підготовку даних з використанням технології NeRF. Також для отримання більш чітких та якісних результатів генерування зображень було введено певні модифікації до оригінальної моделі BARF, а саме: додані функція втрат для оптимізації карт глибин та додатковий вхідний параметр мережі - час t. Остання зміна дозволяє додатково моделювати освітлення сцени, що дозволило б зробити нейромережу для оцінки карт глибин зі стереозображень більш стійкою до змін освітлення.

3 ПРАКТИЧНІ РЕЗУЛЬТАТИ

3.1 Набір даних

Для проведення експериментів було обрано два набора даних: ScanNet [53] та реальний датасет, який був самостійно зібраний за допомогою Samsung Galaxy A22:

1) ScanNet

ScanNet – це набір відеоданих RGB-D, що містить 2,5 мільйона картинок у понад 1500 різних локаціях. Даний датасет належить до indoor наборів даних. Він містить усі параметри камери: внутрішні параметри, зовнішні параметри (пози), а також доступною є семантична сегментація. Основні характеристики камери, що були використані, відображені у табл. 3.1. Пози камери представлені у вигляді матриць, розміром 4 x 4, які відповідають за положення камери для кожного зображення у світовій системі координат. Для проведення усіх експериментів з використанням даного датасету у роботі було обрано 2 сцени: scene0597_00 та scene0000_00. Приклади зображень з даних двох наборів даних проілюстровані на рис. 3.1, 3.2 відповідно.

Таблиця 3.1 – Внутрішні параметри камери для набору даних ScanNet

Фокусна відстань (f_x)	c_x	c_y
577.87	319.5	239.5

2) Самостійно зібраний датасет за допомогою Samsung Galaxy A22

Даний датасет складається зі 123 зображень однієї локації, відзнятої з різних кутів зору. Пози камери є невідомими. Внутрішнні параметри камери були розраховані на основі характеристик сенсору. Отримані значення висвітлені у табл. 3.2. Приклад зображень відображений на рис. 3.3.

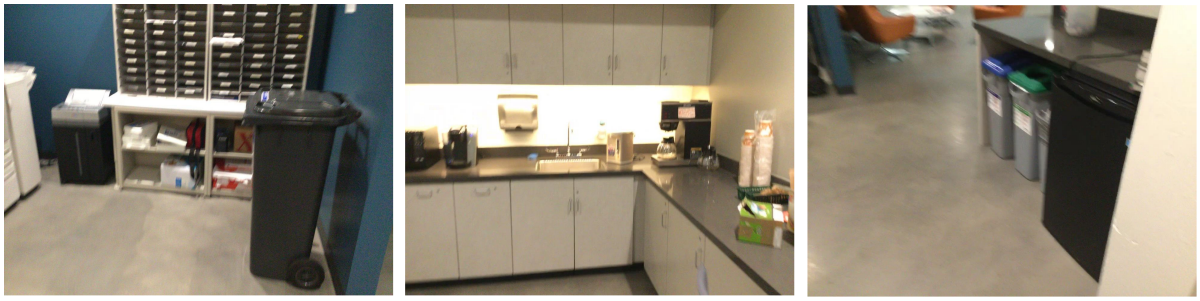


Рисунок 3.1 – Приклад зображень сцени ScanNet scene0597_00

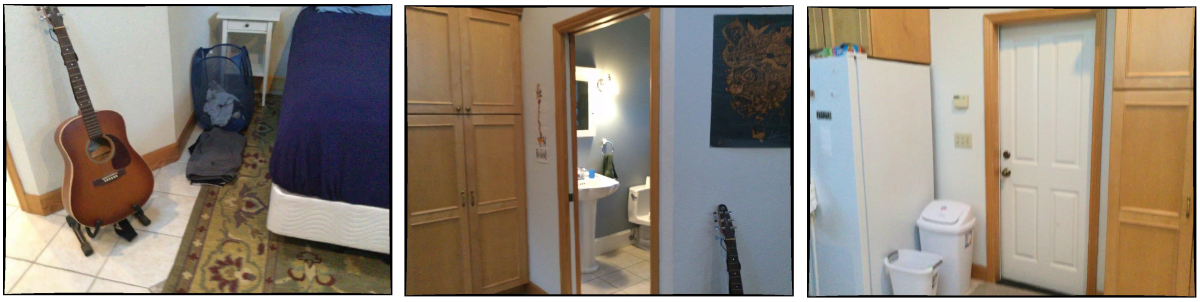


Рисунок 3.2 – Приклад зображень сцени ScanNet scene0000_00

3.2 Генерування стереозображень

Neural Radiance Fields дають змогу генерувати нові зображення, використовуючи будь-які задані параметри камери. Для генерування стереозображень було обрано відстань між лівим та правим зображенням, яка дорівнює 0.09 м. Дана відстань була обрана як приблизне середнє значення між наявними стереобазами у деяких камерах та існуючих датасетах. Наприклад, Intel® RealSense™ Depth Camera D400-Series мають відстані між двома камерами, що дорівнюють 0.05 - 0.055м, а у стереодатасеті Middlebury [9] даний параметр варіюється від 0.14м. Отже, в якості поз камер правих зображень були обрані такі, що є зміщеними по

Таблиця 3.2 – Внутрішні параметри камери для набору даних з Samsung Galaxy A22

Фокусна відстань (f_x)	c_x	c_y
465	320	240

осі x на 0.09м відносно ориганільних поз (лівих) зображень.

3.2.1 Генерування стереозображень без відомих початкових поз камери у наборі даних

Одним з випадків застосування BARF - є генерування даних для сцени, початкові пози камери якої невідомі. Розгляньмо приклад реального датасету, зібраного за допомогою Samsung Galaxy A22 (див. рис. 3.3).

Рисунок 3.3 – Приклад зображень датасету, зібраного за допомогою Samsung Galaxy A22

Так як зовнішні параметри зображень є невідомими, тренування було налаштовано так, що початковими позами для кожного зображення є одиничні матриці, а правильні пози оптимізуються в процессі тренування. Розподіл тренувальних та валідаційних даних був обраний 90/10% відповідно. Приклад згенерованих стереозображень та карти зсувів для лівої фотографії представлений на рис. 3.4.

Для перевірки ефективності генерування стереозображень за допомогою NeRF було також проведено порівняння з іншим методом – Stereo from Mono алгоритмом. Для створення правого зображення були

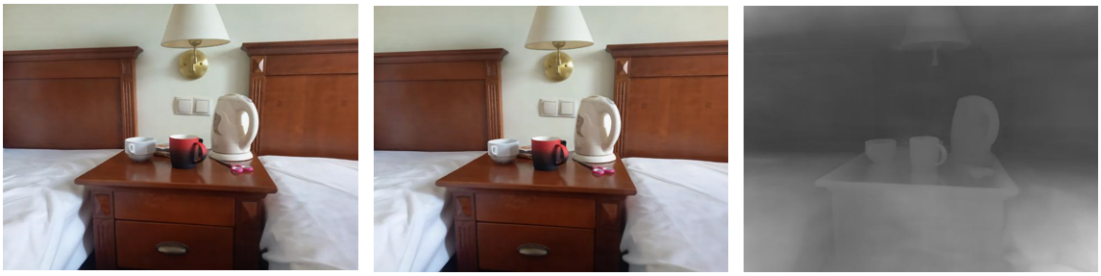


Рисунок 3.4 – Приклад згенерованого лівого (a), правого (b) зображень та карти зсувів для лівої фотографії (c)

обрані ті ж параметри, що і для BARF. Ще одним вхідним параметром алгоритму, окрім початкового зображення та стереобази, є карта зсувів. Для отримання карти глибин було обрано натреновану модель для монокулярної оцінки карт відстаней на основі трансформера – Midas [54]. Вибір моделі у даному випадку не є важливим. Ми обрали Midas адже імплементація мережі є відкритою та є наявні натреновані ваги моделі. Приклад оригінального (лівого) зображення та передбаченої карти глибин представлений на рис. 3.5. Далі, для отримання карти зсувів, була застосована наступна формула [55], використовуючи наявні фокусну відстань (f_x) та стереобазу (baseline):

$$Disp = \frac{f_x * baseline}{depth} \tag{3.1}$$

Використовуючи отриману карту зсувів, ми застосували алгоритм прямого викривлення (forward warping) [15]. Пропущені значення в результаті оклюзій були заповнені за допомогою перенесення кольору [16]. В якості зображення для перенесення кольору було обрано одне із сусідніх, хоча в оригінальному алгоритмі використовується випадково обрана картинка з набору даних. Приклад синтезованого правого зображення та оригінального лівого зображений на рис. 3.6 у другому рядку. У першому рядку представлені, для порівняння, отримана стереопара за допомогою BARF. Результати демонструють, що синтезовані зображення, використовуючи BARF, є більш гладкими, ніж

Рисунок 3.5 – Приклад оригінального лівого зображення та передбаченої карти зсувів за допомогою Midas моделі

оригінальні, проте колір зображень відновлюється правильно та просторові зв'язки є коректними. Згенеровані зображення мають правильно заповнені оклюзійні дірки. Всі деталі об'єктів присутні. У той же час синтезовані зображення за допомогою алгоритму Stereo from Mono мають кращу роздільну здатність, поірівняно з BARF, проте на картинках присутні певні артефакти такі як, наприклад, частини простирадла на комоді або ж пропущені значення, які є некоректно заповненими. Слід зауважити, що передбачені значення відстаней до об'єктів не є ідеально точними. Деталізоване порівняння результатів двох методів представлено на рис. 3.7 у вигляді декількох збільшених частин згенерованих правих зображень.

Якість отриманих стереопар може бути оцінена за допомогою стереонейромереж для оцінки відстаней до об'єктів сцени. Для цього ми вирішили обрати натреновану мережу для оцінки карт глибин зі стереозображень HITNet [56], розроблену Google Research. Передбачені карти зсувів для двох варіантів стереопар, отриманих за допомогою BARF та Stereo from Mono алгоритму, зображено на рис. 3.8. З результатів видно, що карта зсувів для стереопари, синтезованої, використовуючи BARF, має більше структурних деталей (наприклад, торшер, вимикач світла), ніж карта зсувів для стереопари, отриманої за

Рисунок 3.6 – Приклад оригінального лівого зображення та синтезованого правого за допомогою Stereo from Mono алгоритму

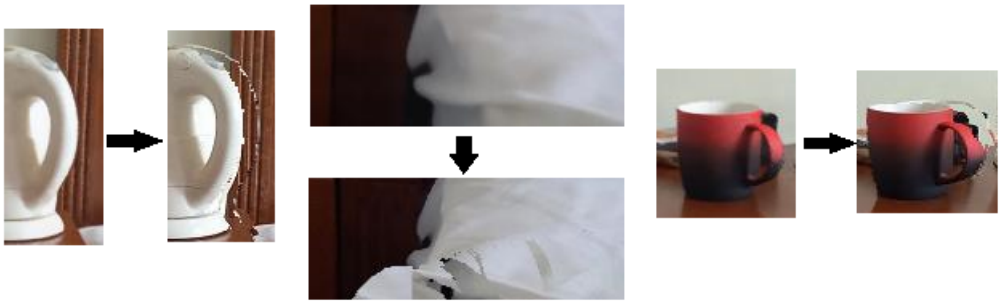


Рисунок 3.7 – Частини зображень, взятих зі згенерованих правих картинок. Ліві зображення відповідають синтезованим NeRF, праві - Stereo from Mono алгоритмом

допомогою іншого алгоритму.

Отже, порівнюючи роботу двох методів – BARF та Stereo from Mono алгоритм, можна зробити такий висновок: перший метод, BARF, показує більш точні результати, на відміну від Stereo from Mono алгоритму, проте має певні недоліки, які виражені у деталізації текстур у згенерованих зображеннях.

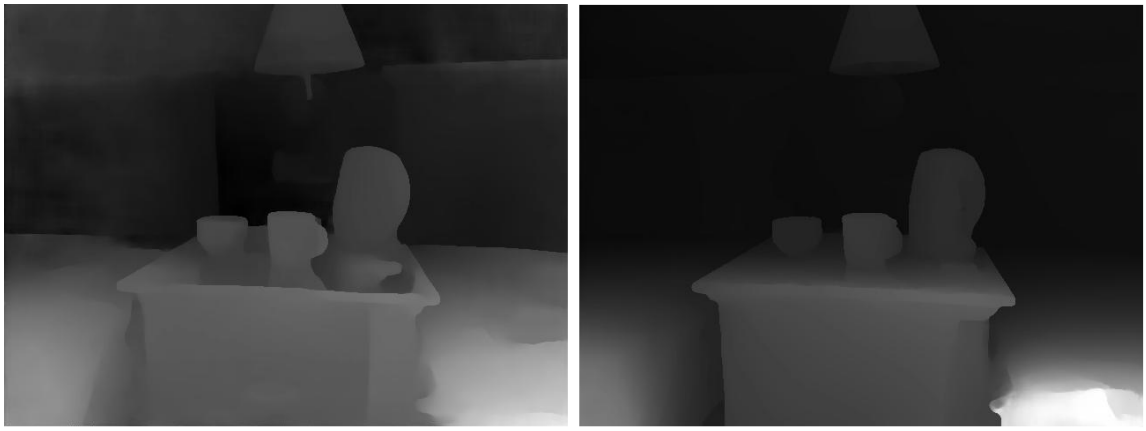


Рисунок 3.8 – Передбачені карти зсувів для стереопар, згенерованих за допомогою BARF (a) та Stereo from Mono алгоритму (b)

3.2.2 Генерування стереозображень з наявними початковими позами камери у наборі даних

У даному підрозділі розгляньмо випадок оригінальної моделі BARF без оптимізації поз камер у процесі тренування. За істинні пози камер було прийнято наявні у датасеті ScanNet для scene0597_00. Приклад синтезованих зображень та відповідних їм карт глибин для зазначеної сцени ScanNet зображений на рис. 3.9. Як видно з картинок, вони є дещо розмитими, а передбачені значення глибин є некоректними.

Для більш детальної оцінки отриманих карт відстаней, синтезовані дані були переведені у 3D у вигляді сітки засобами python бібліотеки open3d [57]. Результати зображені на рис. 3.11. Порівнюючи з істинними даними (рис. 3.10) одразу видно, що структура сцени губиться, наприклад, стіни та інші площини є нерівними.

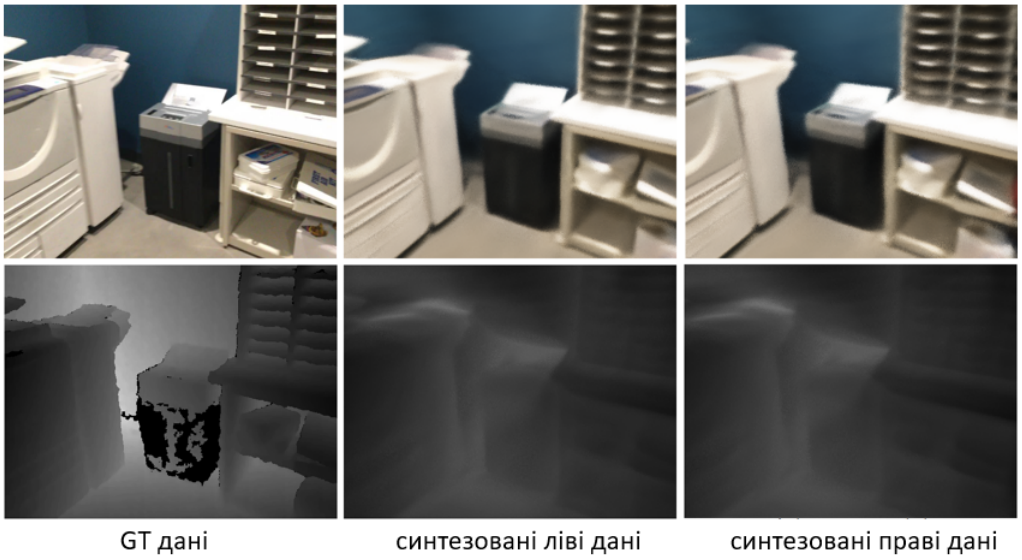


Рисунок 3.9 – Приклад згенерованих стереозображень

Рисунок 3.10 – Приклад GT сітки набору даних scene0597_00

3.3 Додавання нової функції втрат для оцінки карт глибин

У попередньому підрозділі було розглянуто роботу моделі BARF в оригінальному варіанті. У такому випадку передбачені значення глибин

Рисунок 3.11 – Приклад синтезованих оригінальною моделлю BARF карт глибин у вигляді 3D сітки

мають великі розбіжності з істинними даними - відносна помилка досягає ∼14%. Отже, було запропоновано ввести додаткову функцію втрат (див. формула 2.9) для оптимізації передбачених карт глибин.

Щоб перевірити ефективність функції втрат для оцінки карт відстаней, було проведено два тренування: з використанням істинних карт глибин, доступних в ScanNet, та без їх використання. Для тренування була використана одна з тренувальних сцен датасету - scene0597_00. Кількість даних в ній складає 1407 картинок. Тренування було виконано, задіявши кожне друге зображення локації. Приклад роботи мережі представлено на рис. 3.12. З отриманих результатів видно, що відстань до об'єктів передбачається більш коректно, на відміну від отриманих результатів у підрозділі 3.1. Різниця між отриманими відносними помилками для оригінальної моделі BARF та моделі з додатковою функцією втрат сягає ∼12%. Також візуальний аналіз покращень було проведено у 3D просторі, приклади передбачених карт глибин у вигляді 3D моделей зображені на рис. 3.13. Нова функція втрат допомагає подолати проблему нерівності площин. Проте певні артефакти можуть

з'являтися або залишатися без змін, приклади таких недоліків виділені червоними колами на рис. 3.13

Отже, введення додаткової функції втрат, яка мінімізує помилку між істинними картами відстаней та картами відстаней, які передбачає модель BARF, дозволяє покращити отримані результати. Візуальне порівняння передбачених карт глибин двома моделями у вигляді 3D сітки проілюстровано на рис. 3.14.

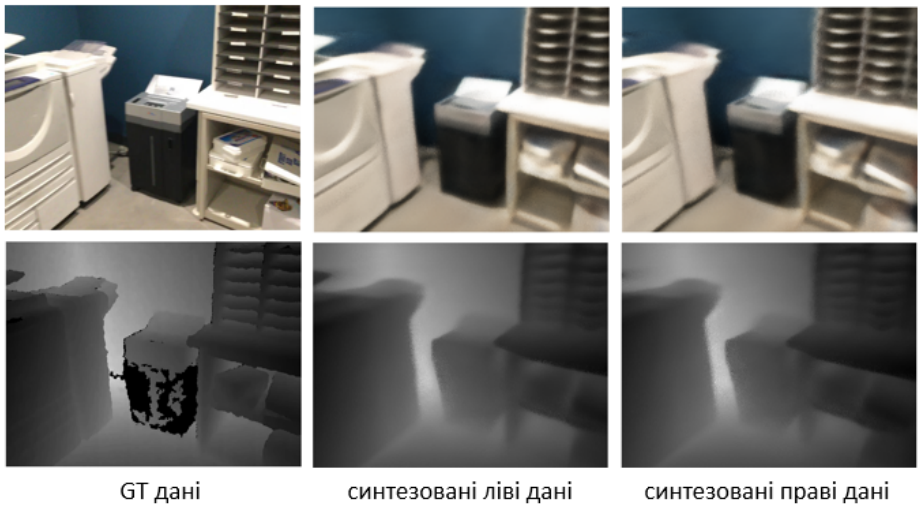


Рисунок 3.12 – Приклад згенерованих стереозображень за допомогою моделі з додатковою функцією втрат для оптимізації карт глибин

3.4 Використання часу t як додаткового параметру моделі BARF

Як було сказано у розділі 2.3.2., до вхідних параметрів моделі було запропоновано додати нову змінну – час t, яка відповідає послідовному індексу картинки, що подається у мережу. Дана зміна може дозволити моделювати освітлення у згенерованих фреймах. Так як очікується, що локації у тренувальному наборі даних є статичними, ми пропонуємо використовувати даний параметр t, щоб вплинути тільки на

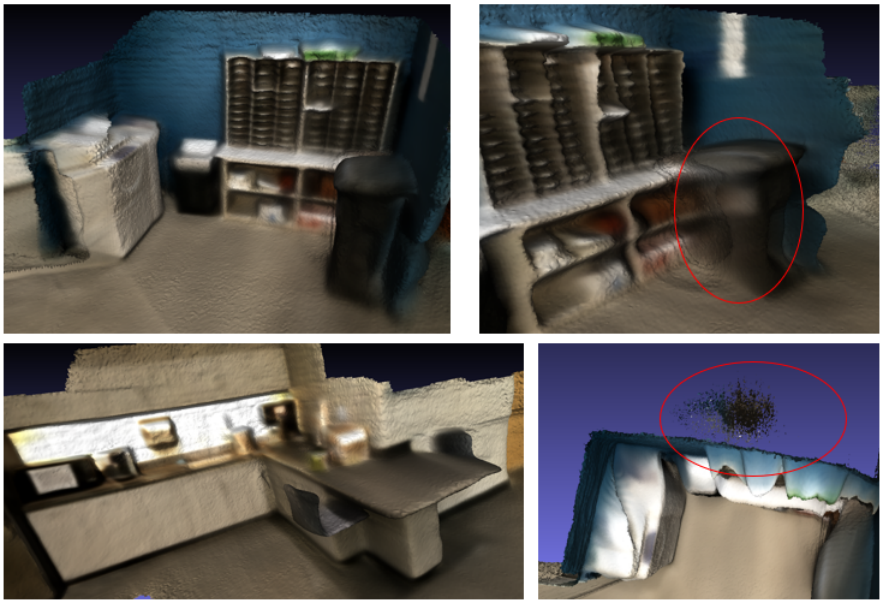


Рисунок 3.13 – Приклад синтезованих BARF карт глибин у вигляді 3D сітки з застосуванням функції втрат для оптимізації карт глибин

результуючий RGB колір. Автори [38] також розглядають введення нового параметру - часу, проте він впливає і на передбачений RGB колір, і на передбачену щільність. Також у зазначеній статті передбачається, що сцена є динамічною. Враховуючи вищесказане, було проведено 2 експерименти з різним використанням нової змінної. Розгляньмо дані два варіанти детальніше.

Видозмінена архітектура BARF разом з додатковою функцією втрат та додатковими вхідними параметрами зображена на рис 3.15. Мережа складається з 8 повнозв'язних шарів, кожен з яких містить 256 нейронів, які на виході передбачають σ та 256-розмірний вектор ознак. Далі отриманий 256-розмірний вектор з'єднується з напрямками вхідних променів та передається до додаткового повнозв'язного шару, який містить 128 нейронів. Останній шар мережі дає змогу отримати RGB колір. Також у моделі присутній зв'язок між вхідними даними та п'ятим повнозв'язним шаром. Таким чином, на вихідну щільність впливають тільки вхідні 3D координати, а на отримуваний колір впливають і координати, і вхідні направлення променів.

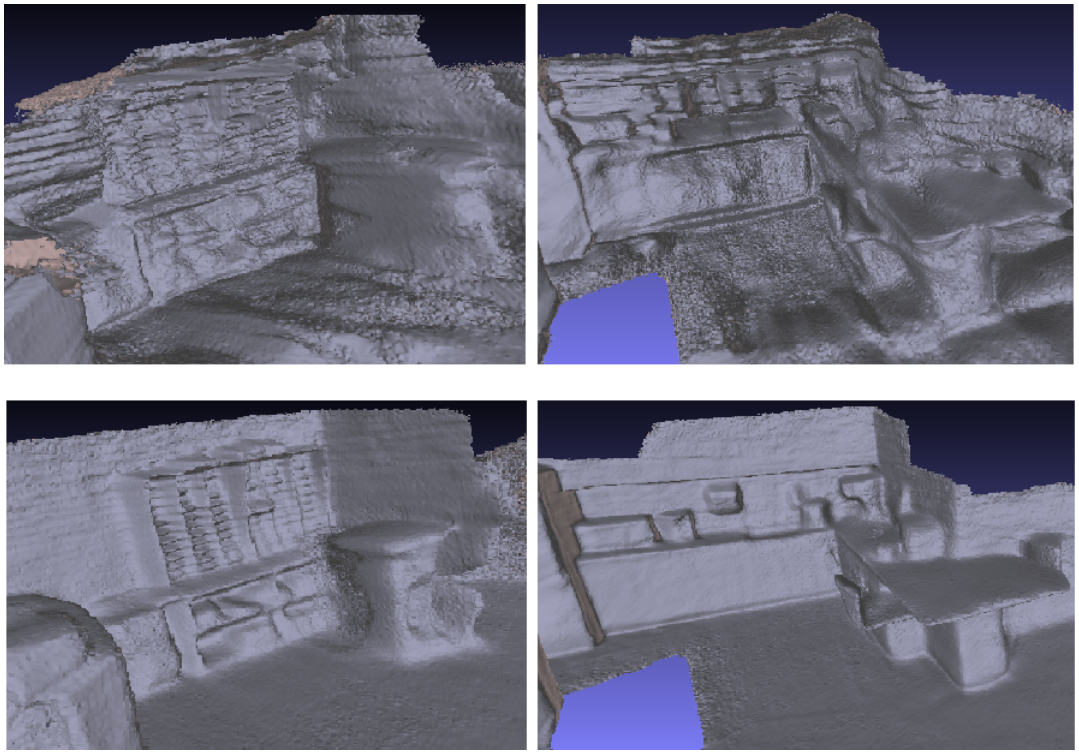


Рисунок 3.14 – Порівняння 3D сіток для оригінальної моделі BARF (верхній рядок) та моделі з введенням функції втрат для оптимізації карт відстаней (нижній рядок)

Для дослідження впливу додаткового вхідного параметру t було проведено 2 експерименти:

1) Введення змінної t впливає тільки на результуючий колір

У архітектурі, яка зображена на рис. 3.15 змінюється вхід Input Direction. Разом з напрямками променів будемо подавати індекс зображення, нормалізований до $[0, 1]$. Тобто вхід виглядає як $(\vec{d}, \vec{t})$.

2) Введення змінної t впливає і на результуючий колір, і на щільність

У архітектурі, яка зображена на рис. 3.15 змінюється вхід Input Ray. Разом з початковими 3D координатами будемо подавати індекс зображення, нормалізований до $[0, 1]$. Тобто вхід виглядає як $(\vec{x}, \vec{t})$.

На рис. 3.16, 3.17 зображені графіки тренувань, які відображають якість роботи кожної з моделей, включаючи варіант тренування, у якому не вводилася додаткова змінна часу t. У відображених результатах було

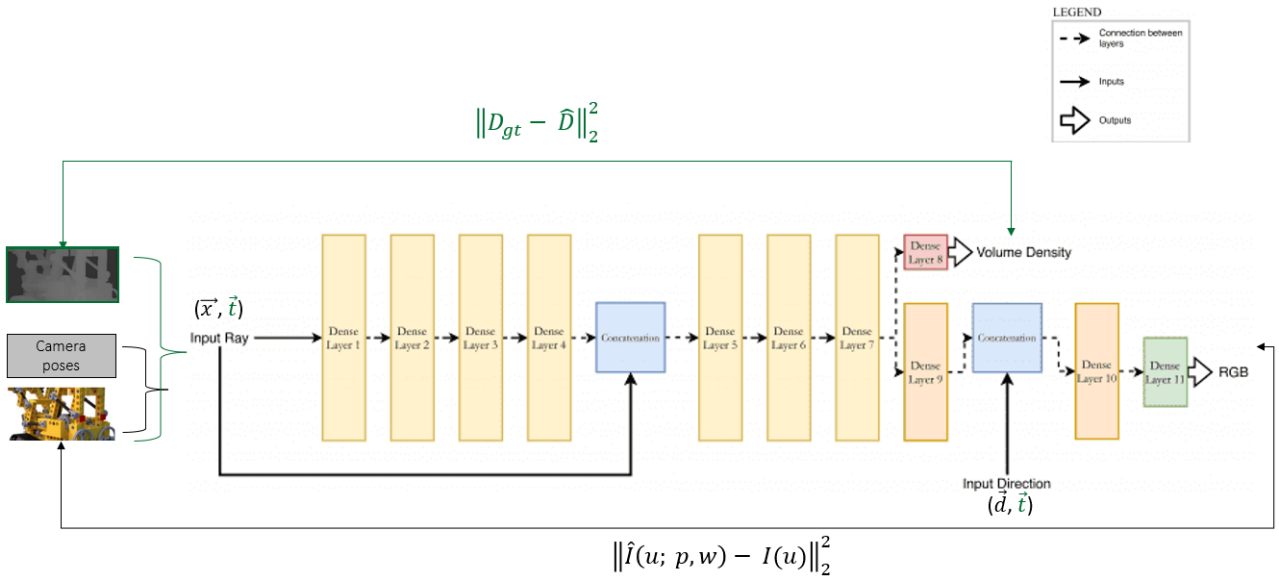


Рисунок 3.15 – Видозмінена архітектура BARF. Зеленим кольором позначені модифікації моделі

використано дві величини: PSNR та помилка синтезу (loss render), що відповідає значенням фотометричних помилок (див. формулу 3.16). Метрика PSNR показує на скільки зображення є спотвореним. У середньому, значення цієї величини, яке дорівнює 40-50 дБ, означає гарний результат.

Означення 3.1. PSNR (Peak Signal-to-Noise Ratio) – співвідношення між максимально можливою потужністю зображення та потужністю спотворюючого шуму, що впливає на якість зображення.

PSNR можна обчислити, використовуючи наступну формулу:

$$PSNR = 10 \log_{10} \frac{(L-1)^2}{MSE} = 20 \log_{10} \frac{L-1}{RMSE} \tag{3.2}$$

де L – максимальна інтенсивність зображення, MSE - середня квадратична помилка між зображеннями, RMSE - квадратний корінь середньої квадратичної помилки

Розгляньмо результуючі PSNR (рис. 3.17) та помилку синтезу (рис. 3.16) детальніше:

1) Введення додаткової вхідної змінної t покращує результати PSNR

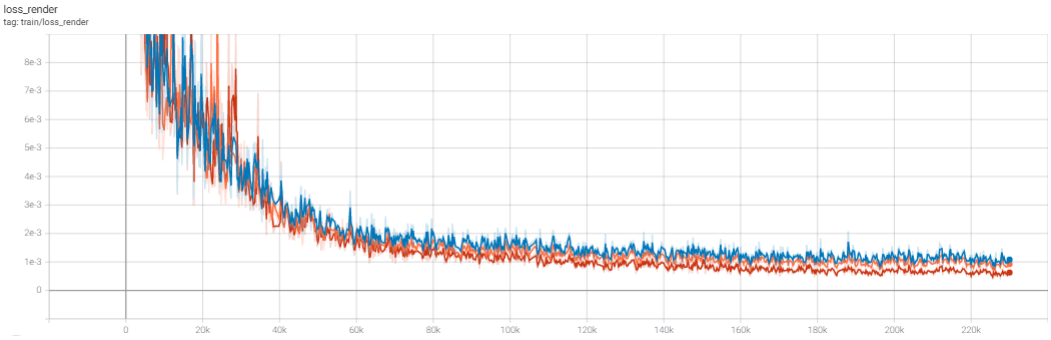


Рисунок 3.16 – Помилки синтезу у тренуваннях без додавання до вхідних параметрів t (синій), з додаванням t разом з напрямками променів (помаранчевий) та з додаванням t разом з 3D координатами (червоний)

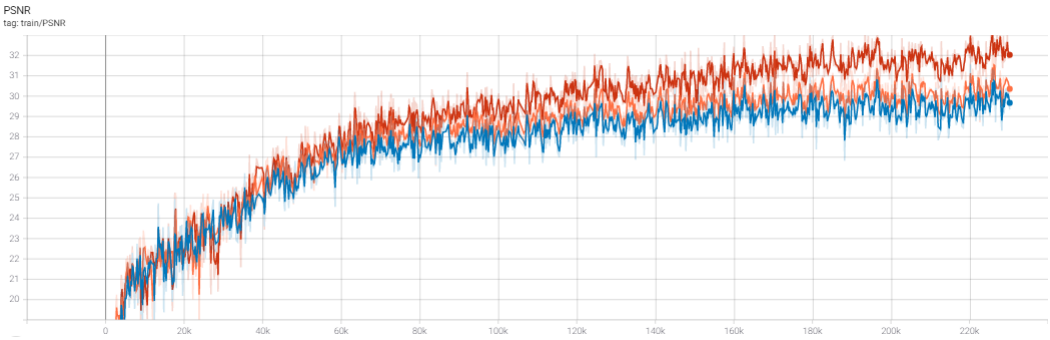


Рисунок 3.17 – PSNR у тренуваннях без додавання до вхідних параметрів t (синій), з додаванням t разом з напрямками променів (помаранчевий) та з додаванням t разом з 3D координатами (червоний)

та loss render у обох випадках

2) У випадку додавання t разом з 3D координатами метрика PSNR досягає значень 30-32 порівняно з тренування без t, яке надає можливість досягнути PSNR значення 29

Результати тренувань у вигляді згенерованих зображень та карт глибин представлені на рис. 3.18, 3.19. Для тренування було використано частину сцени ScanNet scene0000_00. Усього було задіяно 500 картинок. Тренування було здійснено без використання істинних карт глибин. На рис. 3.18, 3.19 червоним кольором позначені певні артефакти такі як, наприклад, шум, а зеленим кольором виділені області, де є покращення даних недоліків.

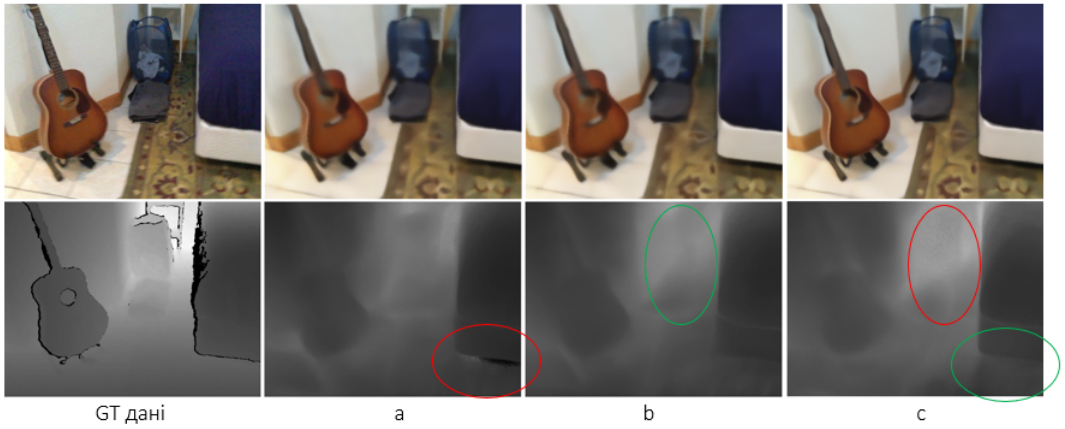


Рисунок 3.18 – Результати тренувань без додавання до вхідних параметрів t (a), з додаванням t разом з напрямками променів (b) та з додаванням t разом з 3D координатами (c)

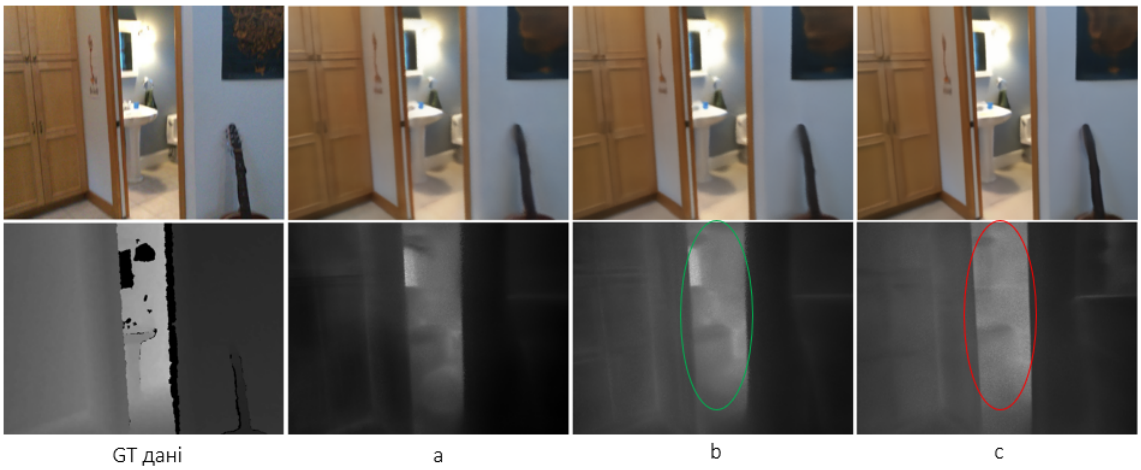


Рисунок 3.19 – Результати тренувань без додавання до вхідних параметрів t (a), з додаванням t разом з напрямками променів (b) та з додаванням t разом з 3D координатами (c)

Отже, з отриманих синтезованих зображень та карт відстаней можна зробити такі висновки:

1) Введення додаткової вхідної змінної t покращує результати синтезу картинок у обох випадках

2) Додатковий параметр t дозволяє отримати більш чіткі та

структуровані карти глибин

3) Додавання параметру t разом з 3D координатами дозволяє отримати більш чіткі результати генерування зображень, ніж у випадку додавання t разом з напрямками вхідних променів

4) У останньому випадку тобто при введенні параметру t разом з 3D координатами у передбачених картах глибин у дальніх областях може з'являтися шум (червоні кола на рис 3.18, 3.19)

Розгляньмо результати тренувань у вигляді 3D сітки. На рис. 3.21 відображений згенерована сітка для трьох випадків: базова модель BARF, модель, яка тренувалася з додаванням t разом з напрямками променів та модель з додаванням параметру t разом з 3D координатами. Червоними колами позначені випадки некоректно згенерованої геометрії, жовтими колами відображені певні покращення відносно червоних, а зеленим - найкращі приклади.

Рисунок 3.20 – GT 3D сітка для scene0000_00

На рис 3.22 відображені ті ж приклади 3D сітки, що і на рис. 3.21, проте також додатково зображений істинна 3D модель. З даних результатів

видно, що з додаванням змінної t структура сцени покращується, проте є невідповідності у масштабі. Вони проявляються у, наприклад, меншому розмірі стін або ж спостерігаються випадки, коли одна зі стін розташована ближче, чим відповідна їй істинна.

Отже, з отриманих результатів можна зробити висновок, що додаткова вхідна змінна t, яка подається у тренування разом з 3D координатами, позитивно впливає на вигляд синтезованих карт глибин, а отже і на 3D реконтрукцію. Проте залишається проблема, пов'язана з невідповідністю масштабів істинних карт глибин та передбачених моделями.

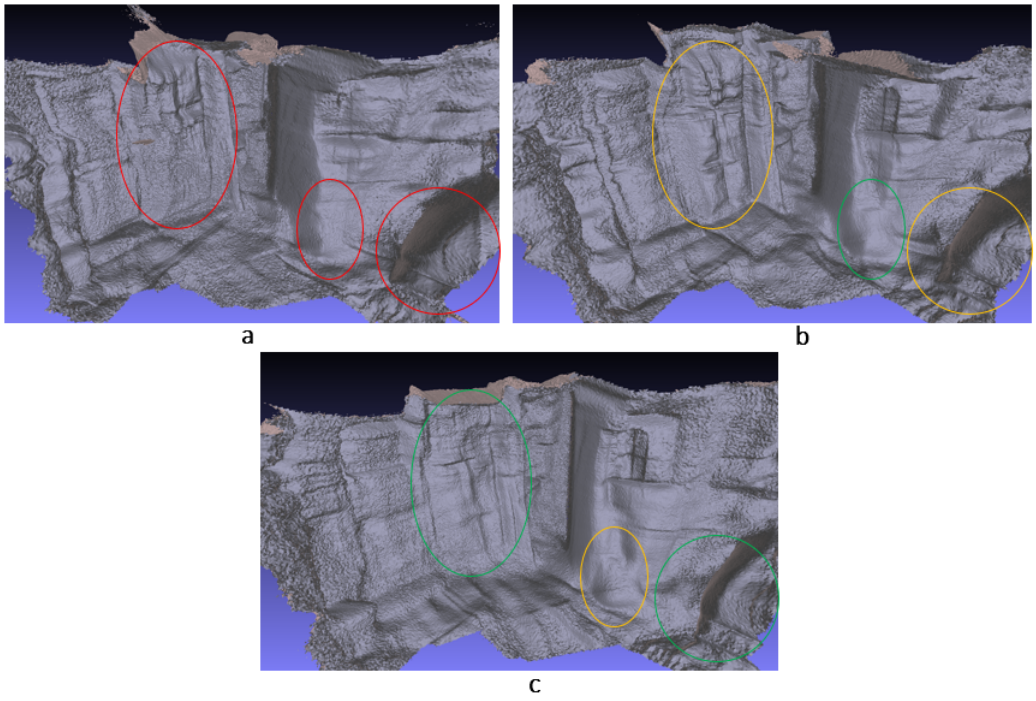


Рисунок 3.21 – Приклади синтезованих карт глибин у вигляді 3D сітки для випадку тренування без додавання t (a), з додаванням t разом з напрямками променів (b) та з додаванням t разом з 3D координатами (c)

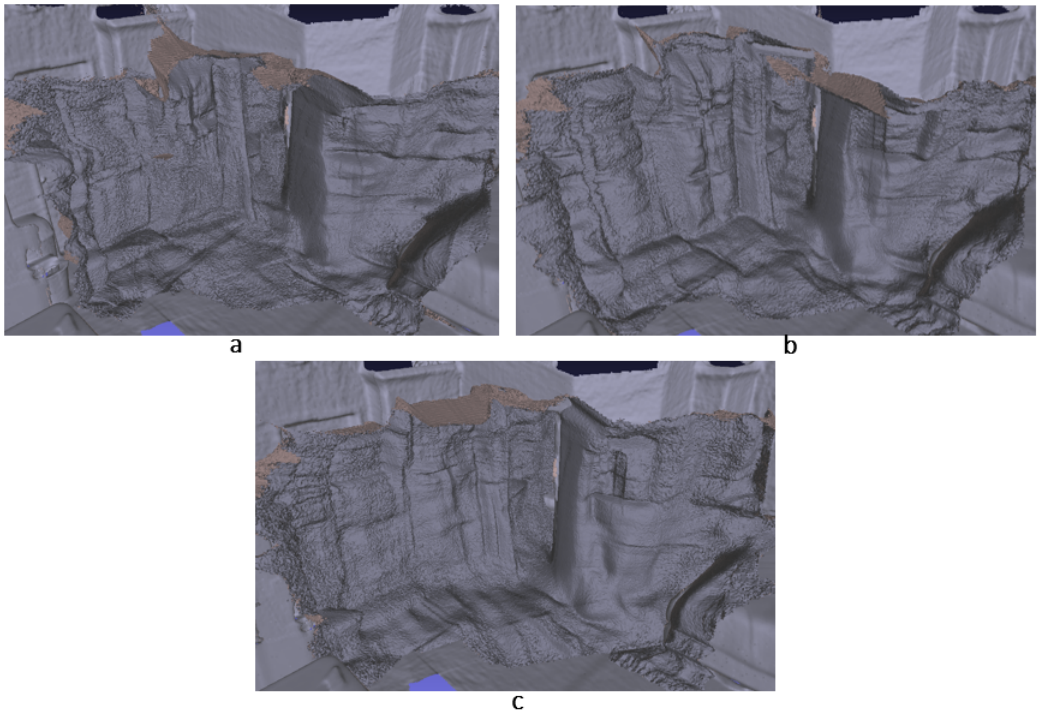


Рисунок 3.22 – Приклади синтезованих карт глибин у вигляді 3D сітки разом з GT 3D моделлю для випадку тренування без додавання t (a), з додаванням t разом з напрямками променів (b) та з додаванням t разом з 3D координатами (c)

3.4.1 Зміна осітлення сцени

Використаймо моделі, отримані у розділі 3.4, для моделювання освітлення. Як вже зазначалося, ідея додавання змінної часу до вхідних параметрів моделі полягає у можливості її використання для зміни освітлення сцени. Тобто постає питання в тому, чи можемо ми, працюючи з датасетом у якому змінюється освітлення з плином часу, змінювати це саме освітлення, фіксуючи певну точку вигляду сцени.

Для проведення експерименту було обрано одне зображення сцени та відповідна йому поза камери. Далі, для генерування нових видів, замість змінної t, яка відповідає номеру фрейма, у мережі подавались інші значення t, які відповідають кадрам датасету з різним освітленням. На рис. 3.23 представлені три оригінальні зображення датасету,

включаючи кадр, позу якого було зафіксовано (300 фрейм). Синтезовані зображення двома моделями з різними варіантами додавання змінної часу проілюстровані на рис. 3.24. Параметр t відповідає індексам фреймів, зображених на рис. 3.23 тобто $t = 300, 357, 465$.

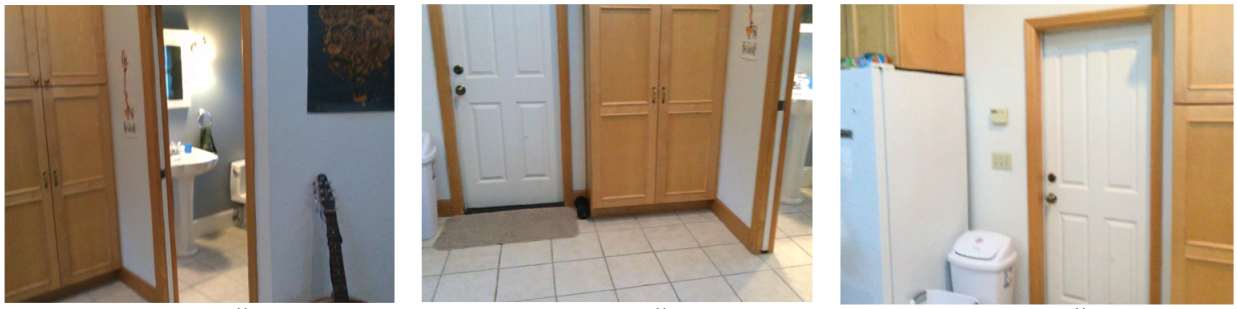


Рисунок 3.23 – Приклади оригінальних зображень з різним освітленням датасету scene0000_00

Рисунок 3.24 – Синтезовані зображення, які моделюють різне освітлення датасету scene0000_00. Верхній рядок відповідає моделі BARF, де змінна часу подається разом з напрямками променів $(\vec{t}, \vec{d})$, а нижній - моделі BARF, де змінна часу подається разом з 3D координатами $(\vec{t}, \vec{x})$

3.5 Аналіз результуючих метрик

Розгляньмо результуючі метрики, які відповідають помилкам у передбачених картах глибин, порівняно з істинними. Для оцінки роботи моделей були використані дві метрики MAE (mean absolute error) та MARE (mean absolute relative error):

MAE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{3.3}$$

де y - істинні карти глибин, $\hat{y}_i$ - передбачені значення глибин, N - кількість зображень

MARE:

$$MARE = \frac{1}{N} * \sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} * 100\% \tag{3.4}$$

де y_i - істинні карти глибин, $\hat{y}_i$ - передбачені значення глибин, N - кількість зображень

У табл. 3.25 відображені результати тренувань у різних режимах у вигляді помилок між GT картами глибин та передбаченими картами глибин для ScanNet сцени 0597, яка була використана у прикладах в підрозділах 3.2 та 3.3. У даному випадку розподіл тренувальних/валідаційний даних був обраний як 95/5%. Розрахунок метрик, наведений у зазначеній таблиці, був проведений на 95% даних. З результатів випливає, що оригінальний варіант BARF показує найгіршу якість генерування даних. У такому випадку вдається досягти помилки в 13.88%, що є досить великим значенням. А найкращим випадком тренування є тренування в режимі додаткової оптимізації карт глибин з додаванням до вхідних параметрів t разом з 3D координатами. Тренування з даними модифікаціями дозволяє отримати помилку 0.8%, що означає, що передбачені значення відстаней до об'єктів є дуже близькими до істинних. Такі значення помилок можуть свідчити про те, що згенеровані карти глибин можуть бути застосовані у тренуванні

стереонейромереж. Також таким чином можна доповнити вже існуючі дані зображеннями з інших кутів зору.

Набір даних	Режим тренування	MAE, м	MARE, %	К-ть зображень для тестування
scene0597_00	w/o GT глибин (оригінальна модель)	0.3264	13.88	1335
scene0597_00	w/o GT глибин + $(\vec{t}, \vec{d})$	0.2893	12.22	1335
scene0597_00	w/o GT глибин + $(\vec{t}, \vec{x})$	0.2585	11.08	1335
scene0597_00	GT глибини	0.0331	1.61	1335
scene0597_00	GT глибини + $(\vec{t}, \vec{d})$	0.02956	1.36	1335
scene0597_00	GT глибини + $(\vec{t}, \vec{x})$	0.0187	0.8	1335
scene0597_00	GT глибини + оптимізація пози	0.0293	1.36	669
scene0597_00	GT глибин + оптимізація пози + $(\vec{t}, \vec{d})$	0.0291	1.33	669

Рисунок 3.25 – Метрики для scene0597_00

У табл. 3.26 наведені значення помилок між істинними значеннями відстаней та передбаченими BARF у різних варіаціях тренувань без використання GT карт глибин для scene0000_00 (рис. 3.20). У оригінальному варіанті BARF без оптимізації поз дозволяє отримати помилку у 28.1%, а найкращий результат - це MARE = 17.3% у режимі з додаванням змінної часу t разом з 3D координатами. Дана сцена має змінне освітлення та помилки для даного датасету є досить високими, проте додавання змінної часу помітно покращує якість 3D реконструкції сцени. Для прикладу, для сцени 0597 такий режим тренування дає меншу помилку - 11%.

Набір даних	Режим тренування	MAE, м	MARE, %	К-ть фреймів для тестування
scene0000_00	w/o GT глибин (оригінальна модель)	0.686	28.1	475
scene0000_00	w/o GT глибин + $(\vec{t}, \vec{d})$	0.498	20	475
scene0000_00	w/o GT глибин + $(\vec{t}, \vec{x})$	0.429	17.3	475

Рисунок 3.26 – Метрики для scene0000_00

3.6 Тренування стереонейромережі для оцінки карт глибин на згенерованих даних

Для оцінки спроможності нейромережі тренуватися на даних, синтезованих за допомогою NeRF, було обрану модель DispNet [13]. Дана мережа має просту архітектуру, схожу до U-net, яка складається з енкодера, за яким слідує декодер. Більш детальний опис наведений у 2 розділі роботи. На вхід моделі подаються 2 зображення – ліве та праве та відповідна карта зсувів. На виході DispNet передбачає карти зсувів між лівими та правими картинками. Було проведено декілька варіантів експериментальних тренувань, використовуючи різні варіації даних, кожен набір, крім одного, складається з 1336 стереопар. Тренування було првоедено, використовуючи ScanNet сцену scene0597_00. Отже, варіанти використаних даних наведені нижче:

1) Істинні дані з датасету.

2) Дані, згенеровані моделлю BARF, яка тренувалася з додатковою оптимізацією карт відстаней.

3) Дані, отримані з моделі, яка тренувалася з оптимізацією карт глибин та додаванням параметра t до входу з 3D координатами.

4) Карти глибин, отримані з моделі, яка тренувалася з оптимізацією карт глибин та додаванням параметра t до входу з 3D координатами. Стереозображення взяті початкові з датасету.

5) Дані, отримані з моделі, яка тренувалася без оптимізації карт глибин та з додаванням параметра t до входу з 3D координатами.

6) Карти глибин, отримані з моделі, яка тренувалася без оптимізації карт глибин та з додаванням параметра t до входу з 3D координатами. Стереозображення взяті початкові з датасету.

7) Дані, згенеровані моделлю BARF, яка тренувалася з додатковою оптимізацією карт відстаней. Також додатково було синтезовано 420 зображень, які відповідають інтерпольованим позам кожної 3 пари

зображень з кроком 3.

У табл. 3.3 наведені результати тренувань DispNet для усіх варіантів, приведених вище.

Окрім експериментів на згенерованих даних, для порівняння були проведені два інших тренування:

1) Тренування DispNet на GT даних датасету scene0597_00 у режимі навчання з учителем.

2) Тренування DispNet на GT даних датасету scene0597_00 у режимі навчання без учителя. Даний режим тренування налаштований, використовуючи метод, описаний у статті [45].

У табл. 3.4 наведені результати тренувань DispNet на істинних даних у режимах з учителем та без учителя, які перелічені вище. Оригінальні дані у датасеті ScanNet є монокулярними, тому для навчання нейромережі стереопари були отримані за допомогою алгоритму Stereo from Mono [14]. Оцінка отриманих результатів, яка представлена в табл. 3.3, 3.4 була здійснена шляхом підрахунку метрик MAE та MARE.

Аналізуючи отримані результати метрик, бачимо, що, використовуючи введені модифікації до моделі BARF, можна приблизити якість тренування стереонейромережі на згенерованих даних до результатів тренувань цієї ж мережі на істинних даних у режимі навчання з учителем. MARE для моделі, яка тренувалася на GT даних, досягає 1.06%, а з модифікаціями - 1.24% - 1.47%. Хоча у такому випадку не вдається отримати кращі результати, важливим моментом є те, що за домогою NeRF можна генерувати будь-яку кількість даних, використовуючи різні параметри каметри. Це є суттєвим обмеженням для таких алгоритмів як, наприклад, Stereo from Mono [14]. Розгляньио результати такого експерименту. У випадку тренування DispNet на синтезованих даних за допомогою моделі BARF з додатковою функцією втрат вдається отримати MARE, яке дорівнює 2.68%. Далі до даного набору даних було додатково згенеровано ще 420 зображень, які відповідають позам камери, котрих немає в оригінальному датасеті.

Таблиця 3.3 – Результати DispNet тренувань, використовуючи згенеровані дані для scene0597_00

Модель BARF	Набір даних	MAE, м	MARE, %	К-ть зобр., датасет	К-ть зобр., тест
w/o GT глибин (оригінальна модель)	scene0597_00	0.280	12.05	1336	1336
GT глибин	scene0597_00	0.057	2.68	1336	1336
GT глибини + $(\vec{t}, \vec{x})$	scene0597_00	0.029	1.47	1336	1336
Істинні зображення + GT глибини + $(\vec{t}, \vec{x})$	scene0597_00	0.225	1.24	1336	1336
w/o GT глибин + $(\vec{t}, \vec{x})$	scene0597_00	0.224	9.81	1336	1336
Істинні зображення + w/o GT глибин + $(\vec{t}, \vec{x})$	scene0597_00	0.251	10.9	1336	1336
GT глибини + 420 нових зображень	scene0597_00	0.040	1.95	1756	1336

Таким чином, тренування стереонейромережі відбувалося на 1756 стереопарах. У такому випадку відносна помилка покращується до 1.95% тобто на 0.7%. Проте слід зазначити, що у даних тренуваннях також спостерігається перетренування так як навчання стереонейромережі відбувалося на одній сцені, яка також входила до валідації.

Отже, проведені експерименти свідчать про те, що стереонейромережу можливо тренувати на згенерованих за допомогою NeRF даних та отримувати близькі результати з навчанням на істинних даних. Проте обмеження використання істинних даних полягяє в їх неузагальненості на різні параметри камери. Проведений експеримент з догенеруванням даних показує, що NeRF може обходити дане обмеження та така аугментація потенційно може покращувати результуючі метрики.

Таблиця 3.4 – Результати DispNet тренувань, використовуючи істинні дані scene0597_00

Модель	Набір даних	MAE, м	MARE, %	К-ть зобр., датасет	К-ть зобр., тест
GT дані, навчання з учителем	scene0597_00	0.022	1.06	1336	1336
GT дані, навчання без учителя	scene0597_00	0.218	8.57	1336	1336

3.7 Додаткові можливі use-cases

Більшість наборів даних для тренування нейронних мереж для оцінки карт глибин мають пропущені значення у своїх істинних даних. Одна з причин -– це певні обмеження сенсорів. Одним з можливих варіантів використання NeRF є заповнення істинних карт глибин згенерованими зображеннями за допомогою технології NeRF.

Для генерування таких зображень модель NeRF була попередньо натренована на обраній сцені датасету ScanNet - .scene0597_00. Далі замість пропущенних значень у істинних картах відстаней були використані передбачені NeRF. Приклади результату заповнення нульових значень глибин зображено на рис. 3.27.

Також приклади результатів окремих карт відстаней в 3D з та без доповнення карт глибин зображено на рис. 3.28, а частина 3D сітки усієї сцени з заповненням пропущених значень за доп. NeRF зображена на рис. 3.29. Отже, з отриманих результатів можна зробити такі висновки:

1) Пропущені значення в істинних картах глибин можуть бути заповнені, використовуючи передбачені значення NeRF.

2) Порівнюючи на прикладі набора даних ScanNet індивідуальні карти відстаней у 3D для істинних даних та доповненних, другі суттєво покращують вигляд сітки та мають коректну геометрію у нульових зонах

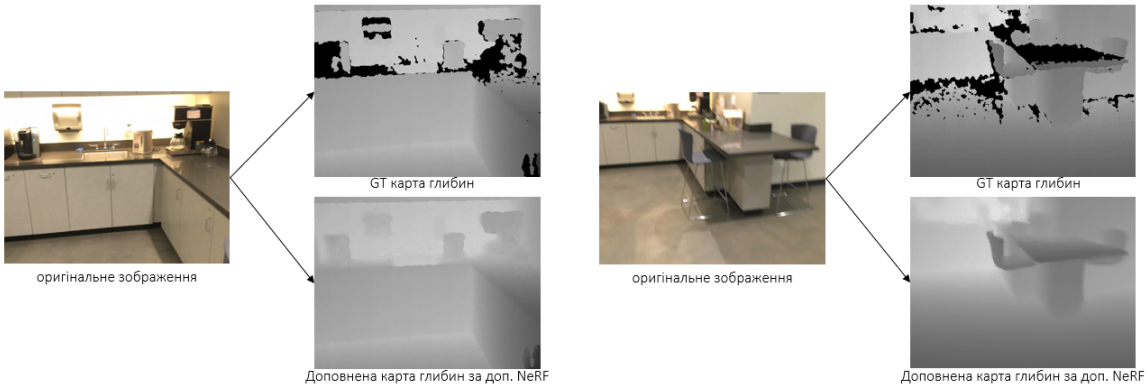


Рисунок 3.27 – Приклад доповнення карт глибин синтезованими значеннями за допопомгою BARF

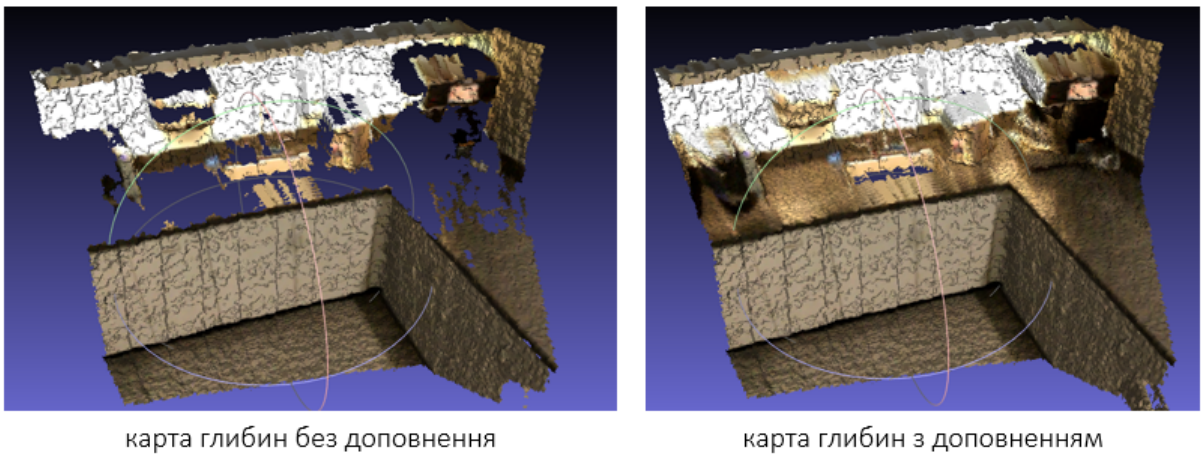


Рисунок 3.28 – Приклад індивідуальних карт глибин в 3D

(рис. 3.28, 3.29)

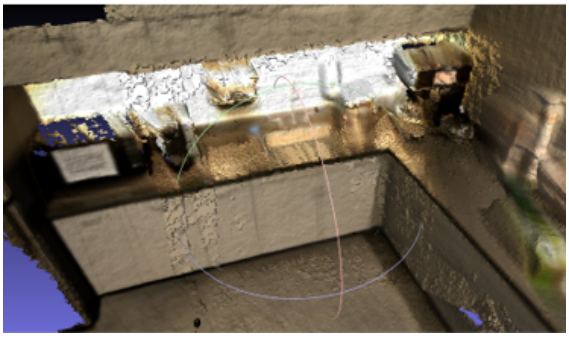


Рисунок 3.29 – Приклад частини 3D сітки усієї сцени з заповненням пропущених значень за доп. NeRF

Висновки до розділу 3

У даному розділі були відображені результати проведених експериментів. Основними висновками є те, що видозмінений варіант моделі BARF дає змогу генерувати дані, які мають помилку з істинними $< 1\%$. Також додатковий параметр часу відіграє суттєву роль у отриманій якості згенерованих даних особливо у випадку сцен зі змінним освітленням, зменшуючи MARE більше, ніж на 10%, порівняно з результатами оригінальної моделі.

Слід зазначити, що згенеровані навчальні дані за допопомогою NeRF не покращують результати роботи стереонейромережі, порівняно з навчанням на істинних даних. Проте технологія NeRF дає змогу генерувати нові зображення, обходячи обмеження GT даних щодо неузагальненості на інші параметри камери. У одному з експериментім було показано, що розширення набору даних додатково синтезованими даними, використовуючи NeRF, дає змогу покращити результати роботи стереонейромережі.

ВИСНОВКИ

Головним завданням даної магістерської дисертації було проаналізувати можливість генерування даних для задач стереозору за допомогою технології NeRF. В якості моделі для аналізу було обрано BARF.

У процесі роботи було оглянуто основні видозмінені моделі NeRF та один з алгоритмів, який може бути використаний для генерування стереопар з будь-якої послідовності монокулярних зображень. Проте останній алгоритм має свої недоліки та обмеження. Було встановлено, що у літературі не досліджено можливість використання NeRF для генерування стереоданих.

За результатами експериментів було виявлено, що оригінальна модель BARF не дає можливості отримати дані високої якості. Отже, у даній роботі було запропоновано ввести дві модифікації для покращення генерування стереопар:

1) Додаткова функція втрат для оптимізації передбачених карт глибин.

2) Додатковий вхідний параметр моделі час t, який відповідає послідовному індексу вхідного зображення.

За результатами аналізу проведених експериментів було встановлено:

1) Додаткова функція втрат для оптимізації карт глибин дозволяє покращити результат роботи BARF на $\sim$12%.

2) Додатковий вхідний параметр часу t покращує результат роботи BARF, проте покращення залежить від набору даних та освітлення. У одному випадку це 1-2%, в іншому – 11%.

3) Об'єднання двох модифікацій і застосування їх разом дозволяє отримати помилку $< 1\%$.

4) Додатковий вхідний параметр часу t надає змогу моделювати освітлення сцени.

5) Застосовуючи запропоновані модифікації до генерування даних, можна наблизити результати роботи стереонейромережі на згенерованих даних до результатів роботи стереонейромережі на істинних даних (1.47%/1.24% (BARF + модифікації) -> 1.06% (GT дані)).

6) NeRF дає змогу генерувати нові зображення сцени, використовуючи різні параметри камери. Аугментація таким чином навчальних даних дозволяє покращити результат роботи стереонейромережі на 0.7%.

Отримані результати дають змогу синтезувати стереозображення з наявних монокулярних наборів даних та застосовувати їх для покращення якості результатів навчання нейромереж для оцінки карт глибин зі стереозображень.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. On the confidence of stereo matching in a deep-learning era: a quantitative evaluation / Poggi Matteo, Kim Seungryong, Tosi Fabio, Kim Sunok, Aleotti Filippo, Min Dongbo, Sohn Kwanghoon, and Mattoccia Stefano // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2021. — 04. — Vol. PP. — P. 1–1.

2. Tardon Lorenzo, Barbancho Isabel, Alberola-López Carlos. Markov Random Fields in the Context of Stereo Vision. — 2011. — 01. — ISBN: 978-953-307-516-7.

3. Zhu Shiping, Yan Lina. Local Stereo Matching Algorithm with Efficient Matching Cost and Adaptive Guided Image Filter. — 2017. — Vol. 33, no. 9. — Access mode: https://doi.org/10.1007/s00371-016-1264-6.

4. Bleyer Michael, Rhemann Christoph, Rother Carsten. PatchMatch Stereo - Stereo Matching with Slanted Support Windows // BMVC. — 2011. — January. — Access mode: https://www.microsoft.com/en-us/research/publication/patchmatch-stereo-stereo-matching-with-slanted-support-windows/.

5. A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation / Laga Hamid, Jospin Laurent Valentin, Boussaid Farid, and Bennamoun Mohammed // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2022. — apr. — Vol. 44, no. 4. — P. 1738–1764.

6. Chapter 4 - Multiview HDR Video Sequence Generation / Orozco R.R., Loscos C., Martin I., and Artusi A. // High Dynamic Range Video / ed. by Dufaux Frédéric, Le Callet Patrick, Mantiuk Rafał K., Mrak Marta. — Academic Press, 2016. — P. 121–138. — Access mode: https://www.sciencedirect.com/science/article/pii/B9780081004128000048.

7. Menze Moritz, Geiger Andreas. Object Scene Flow for Autonomous

Vehicles // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2015. — June.

8. DrivingStereo: A Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios / Yang Guorun, Song Xiao, Huang Chaoqin, Deng Zhidong, Shi Jianping, and Zhou Bolei // IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2019.

9. High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth. / Scharstein Daniel, Hirschmüller Heiko, Kitajima York, Krathwohl Greg, Nesic Nera, Wang Xi, and Westling Porter // GCPR / ed. by Jiang Xiaoyi, Hornegger Joachim, Koch Reinhard. — Springer. — 2014. — Vol. 8753 of Lecture Notes in Computer Science. — P. 31–42. — Access mode: http://dblp.uni-trier.de/db/conf/dagm/gcpr2014.html#ScharsteinHKKNWW14.

10. A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Videos / Schops Thomas, Schonberger Johannes L., Galliani Silvano, Sattler Torsten, Schindler Konrad, Pollefeys Marc, and Geiger Andreas // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2017. — July.

11. VirtualWorlds as Proxy for Multi-object Tracking Analysis / Gaidon Adrien, Wang Qiao, Cabon Yohann, and Vig Eleonora // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 2016. — P. 4340–4349.

12. Xiao Jianxiong, Owens Andrew, Torralba Antonio. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels // 2013 IEEE International Conference on Computer Vision. — 2013. — P. 1625–1632.

13. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation / Mayer Nikolaus, Ilg Eddy, Hausser Philip, Fischer Philipp, Cremers Daniel, Dosovitskiy Alexey, and Brox Thomas // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — IEEE. — 2016. — jun.

14. Learning Stereo from Single Images / Watson Jamie, Aodha Oisin Mac, Turmukhambetov Daniyar, Brostow Gabriel J., and Firman Michael. — Berlin, Heidelberg : Springer-Verlag. — 2020. — Access mode: https://doi.org/10.1007/978-3-030-58452-8_42.

15. Schwarz L.A. Non-rigid registration using free-form deformations : Ph. D. thesis. — Technische Universitat Munchen, 2007.

16. Color transfer between images / Reinhard E., Adhikhmin M., Gooch B., and Shirley P. // IEEE Computer Graphics and Applications. — 2001. — Vol. 21, no. 5. — P. 34–41.

17. Waechter Michael, Moehrle Nils, Goesele Michael. Let There Be Color! Large-Scale Texturing of 3D Reconstructions // Computer Vision – ECCV 2014 / ed. by Fleet David, Pajdla Tomas, Schiele Bernt, Tuytelaars Tinne. — Cham : Springer International Publishing. — 2014. — P. 836–850.

18. Local Implicit Grid Representations for 3D Scenes. — 2020. — Access mode: https://arxiv.org/abs/2003.08981.

19. Penner Eric, Zhang Li. Soft 3D Reconstruction for View Synthesis. — 2017. — Vol. 36, no. 6.

20. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis / Mildenhall Ben, Srinivasan Pratul P., Tancik Matthew, Barron Jonathan T., Ramamoorthi Ravi, and Ng Ren // ECCV. — 2020.

21. Kajiya James T., Herzen Brian Von. Ray tracing volume densities // Proceedings of the 11th annual conference on Computer graphics and interactive techniques. — 1984.

22. Learning Object-Compositional Neural Radiance Field for Editable Scene Rendering / Yang Bangbang, Zhang Yinda, Xu Yinghao, Li Yijin, Zhou Han, Bao Hujun, Zhang Guofeng, and Cui Zhaopeng // International Conference on Computer Vision (ICCV). — 2021. — October.

23. Editable Free-Viewpoint Video using a Layered Neural Representation / Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang,

Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu // ACM SIGGRAPH. — 2021.

24. Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation. — 2022. — Access mode: https://arxiv.org/abs/2205.04334.

25. In-Place Scene Labelling and Understanding with Implicit Scene Representation / Zhi Shuaifeng, Laidlow Tristan, Leutenegger Stefan, and Davison Andrew // Proceedings of the International Conference on Computer Vision (ICCV). — 2021.

26. NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo / Wei Yi, Liu Shaohui, Rao Yongming, Zhao Wang, Lu Jiwen, and Zhou Jie // ICCV. — 2021.

27. MINE: Towards Continuous Depth MPI with NeRF for Novel View Synthesis / Li Jiaxin, Feng Zijian, She Qi, Ding Henghui, Wang Changhu, and Lee Gim Hee // ICCV. — 2021.

28. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. — 2021. — Access mode: https://arxiv.org/abs/2106.10689.

29. NICE-SLAM: Neural Implicit Scalable Encoding for SLAM / Zhu Zihan, Peng Songyou, Larsson Viktor, Xu Weiwei, Bao Hujun, Cui Zhaopeng, Oswald Martin R., and Pollefeys Marc // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2022. — June.

30. BARF: Bundle-Adjusting Neural Radiance Fields / Lin Chen-Hsuan, Ma Wei-Chiu, Torralba Antonio, and Lucey Simon // IEEE International Conference on Computer Vision (ICCV). — 2021.

31. Schönberger Johannes Lutz, Frahm Jan-Michael. Structure-from-Motion Revisited // Conference on Computer Vision and Pattern Recognition (CVPR). — 2016.

32. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. — 2020. — Access mode: https://arxiv.org/abs/2006.10739.

33. NerfingMVS: Guided Optimization of Neural Radiance Fields for Indoor Multi-view Stereo. — 2021. — Access mode: https://arxiv.org/abs/2109.01129.

34. Dense Depth Priors for Neural Radiance Fields from Sparse Input Views / Roessle Barbara, Barron Jonathan T., Mildenhall Ben, Srinivasan Pratul P., and Nießner Matthias // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2022. — June.

35. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo / Chen Anpei, Xu Zexiang, Zhao Fuqiang, Zhang Xiaoshuai, Xiang Fanbo, Yu Jingyi, and Su Hao // Proceedings of the IEEE/CVF International Conference on Computer Vision. — 2021. — P. 14124–14133.

36. Large Scale Multi-view Stereopsis Evaluation / Jensen Rasmus, Dahl Anders, Vogiatzis George, Tola Engil, and Aanæs Henrik // 2014 IEEE Conference on Computer Vision and Pattern Recognition. — 2014. — P. 406–413.

37. iMAP: Implicit Mapping and Positioning in Real-Time / Sucar Edgar, Liu Shikun, Ortiz Joseph, and Davison Andrew // Proceedings of the International Conference on Computer Vision (ICCV). — 2021.

38. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes. — 2020. — Access mode: https://arxiv.org/abs/2011.13084.

39. LENS: Localization enhanced by NeRF synthesis. — 2021. — Access mode: https://arxiv.org/abs/2110.06558.

40. Kendall Alex, Grimes Matthew, Cipolla Roberto. PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization. — 2015. — Access mode: https://arxiv.org/abs/1505.07427.

41. Real-Time RGB-D Camera Relocalization / Glocker Ben, Izadi Shahram, Shotton Jamie, and Criminisi Antonio // International Symposium on Mixed and Augmented Reality (ISMAR). — IEEE. — 2013. — October. — Access mode: https://www.microsoft.com/en-us/research/publication/real-time-rgb-d-camera-relocalization/.

42. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections / Martin-Brualla Ricardo, Radwan Noha, Sajjadi Mehdi S. M., Barron Jonathan T., and Dosovitskiy Alexey Duckworth Daniel // CVPR. — 2021.

43. Godard Clément, Mac Aodha Oisin, Brostow Gabriel J. Unsupervised Monocular Depth Estimation with Left-Right Consistency. — 2016. — Access mode: https://arxiv.org/abs/1609.03677.

44. Smolyanskiy Nikolai, Kamenev Alexey, Birchfield Stan. On the Importance of Stereo for Accurate Depth Estimation: An Efficient Semi-Supervised Deep Neural Network Approach. — 2018. — Access mode: https://arxiv.org/abs/1803.09719.

45. Unsupervised Monocular Depth Learning in Dynamic Scenes. — 2020. — Access mode: https://arxiv.org/abs/2010.16404.

46. Progressive Fusion for Unsupervised Binocular Depth Estimation using Cycled Networks. — 2019. — Access mode: https://arxiv.org/abs/1909.07667.

47. Ronneberger Olaf, Fischer Philipp, Brox Thomas. U-Net: Convolutional Networks for Biomedical Image Segmentation. — 2015. — Access mode: https://arxiv.org/abs/1505.04597.

48. End-to-End Learning of Geometry and Context for Deep Stereo Regression. — 2017. — Access mode: https://arxiv.org/abs/1703.04309.

49. Cascade Residual Learning: A Two-stage Convolutional Neural Network for Stereo Matching. — 2017. — Access mode: https://arxiv.org/abs/1708.09204.

50. A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation / Laga Hamid, Jospin Laurent Valentin, Boussaid Farid, and Bennamoun Mohammed // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2022. — apr. — Vol. 44, no. 4. — P. 1738–1764. — Access mode: https://doi.org/10.1109%2Ftpami.2020.3032602.

51. Regularizing Nighttime Weirdness: Efficient Self-supervised Monocular Depth Estimation in the Dark. — 2021. — Access mode: https://arxiv.org/abs/2108.03830.

52. Digging Into Self-Supervised Monocular Depth Estimation. — 2018. — Access mode: https://arxiv.org/abs/1806.01260.

53. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes / Dai Angela, Chang Angel X., Savva Manolis, Halber Maciej, Funkhouser Thomas, and Nießner Matthias // Proc. Computer Vision and Pattern Recognition (CVPR), IEEE. — 2017.

54. Ranftl René, Bochkovskiy Alexey, Koltun Vladlen. Vision Transformers for Dense Prediction // ICCV. — 2021.

55. Lucas Bruce, Kanade Takeo. An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI). — 1981. — 04. — Vol. 81.

56. HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. — 2020. — Access mode: https://arxiv.org/abs/2007.12140.

57. Zhou Qian-Yi, Park Jaesik, Koltun Vladlen. Open3D: A Modern Library for 3D Data Processing. — 2018. — Access mode: https://arxiv.org/abs/1801.09847.