Міністерство освіти і науки України
Львівський національний університет імені Івана Франка
Факультет електроніки та комп’ютерних технологій
Кафедра оптоелектроніки та інформаційних технологій

Курсова робота

Визначення відстані до об’єкта на зображенні за методами глибинного навчання

Виконав:
студент групи ФеіМ-11
спеціальності 122 –
Комп’ютерні науки:
Олександр РАДЬ

Науковий керівник:
доц. Олексій КУШНІР

Львів 2025

Анотація

У цій курсовій роботі досліджується методика визначення відстані до об’єктів на зображенні з використанням методів глибинного навчання. Основною метою є розробка та реалізація підходу для оцінки відстані до об’єктів на основі стереопари зображень. Процес включає застосування сучасних технік комп’ютерного зору для виявлення об’єктів та вилучення їхніх ознак. Для підтвердження відповідності об’єктів на обох зображеннях використовується згорткова нейронна мережа. Фінальний розрахунок відстані здійснюється шляхом аналізу відмінностей у положенні центрів відповідних об'єктів на стереозображеннях.

Робота також висвітлює можливості застосування отриманих результатів у різних сферах, таких як: автономні транспортні засоби, робототехніка, доповнена та віртуальна реальність.

Ключові слова: глибинне навчання, стереопара, обчислення відстані, комп'ютерний зір.

Abstract

This term paper investigates the methodology for determining the distance to objects in an image using deep learning methods. The main goal is to develop and implement an approach for estimating the distance to objects based on a stereo image pair. The process includes the use of modern computer vision techniques to detect objects and extract their features. A convolutional neural network is used to confirm the correspondence of objects in both images. The final distance calculation is performed by analyzing the differences in the position of the centers of the corresponding objects in the stereo images.

The paper also highlights the possibilities of applying the obtained results in various fields, such as autonomous vehicles, robotics, augmented and virtual reality.

Keywords: deep learning, stereo pair, distance calculation, computer vision.

Зміст

Перелік умовних позначень

1. CNN (Convolutional Neural Network) – нейронна мережа, призначена для обробки та аналізу великих обсягів зображення.
2. Базова лінія (baseline, *B*) – відстань між двома камерами стереопари.
3. Диспаратність – видиме зміщення положення об’єкта між лівим та правим зображеннями стереопари.
4. Ректифікація – процес геометричного перетворення зображень стереопари таким чином, щоб відповідні точки на обох зображеннях лежали на одних і тих же горизонтальних лініях сканування.
5. Бінокулярна диспаратність (далі – диспаратність) – це різниця взаємного положення об’єкта, який спостерігався лівим і правим оком, що є наслідком горизонтальної сепарації очей.

Вступ

Сприйняття та розуміння тривимірного простору — це базова властивість людини, без якої неможливо орієнтуватися у світі, взаємодіяти з об'єктами або оцінювати відстані. Відтворення цієї здатності у машинах є одним із ключових завдань сучасної науки та інженерії. З огляду на активний розвиток автономних систем, робототехніки, технологій доповненої та віртуальної реальності, а також систем безпеки й моніторингу, наділення машин можливістю «бачити» та аналізувати простір набуває особливої актуальності.

Комп'ютерний зір – це галузь штучного інтелекту, яка прагне надати комп'ютерам можливість «бачити» та інтерпретувати зображення або відео так само, як це робить людина. Основна мета комп'ютерного зору полягає у розробці теорій та алгоритмів, що дозволяють автоматизованим системам витягувати, обробляти, аналізувати та розуміти інформацію з цифрових зображень. Типові завдання комп'ютерного зору включають розпізнавання об'єктів, класифікацію зображень, сегментацію, відстеження руху, а також реконструкцію тривимірного простору на основі двовимірних зображень.

Традиційно, завдання визначення відстані в комп'ютерному зорі вирішувалося різними методами, зокрема, із застосуванням стереозору (аналізу двох зображень, знятих з відомої відстані), структури з руху (аналізу послідовності зображень з рухомої камери) або використання активних датчиків (лідари, радари, структурне освітлення). Кожен з цих підходів має свої переваги та обмеження. Метод стереозору, імітуючи бінокулярний зір людини, є одним із найпоширеніших пасивних методів оцінки глибини, заснований на принципі триангуляції та аналізі диспаратності – зміщення відповідних точок або об'єктів на лівому та правому зображеннях стереопари.

За останні роки, людство досягло значного прориву в галузі глибинного навчання, що застосовує багатошарові нейронні мережі для автоматизованого

виявлення складних закономірностей у даних без необхідності ручного виділення ознак. Особливо помітні досягнення демонструють згорткові нейронні мережі (CNN), які відзначаються високою ефективністю в обробці зображень, досягаючи значних успіхів у задачах класифікації та розпізнавання об’єктів. Інтеграція методів глибинного навчання у комп’ютерний зір суттєво підвищила точність і надійність різноманітних процесів, зокрема виявлення об’єктів та оцінки глибини.

Обчислення відстані з зображень є критично важливим етапом для багатьох застосувань. У контексті автономних систем, точна інформація про відстань до інших транспортних засобів, пішоходів чи перешкод є життєво необхідною для безпечної навігації та прийняття рішень. У робототехніці дані про глибину дозволяють роботу маніпулювати об’єктами та орієнтуватися в просторі. Для систем доповненої реальності точне розміщення віртуальних об’єктів у реальному світі неможливе без розуміння геометрії сцени та відстаней до її елементів.

Незважаючи на досягнення у цій сфері, точне визначення відстані до конкретного об’єкта на зображенні, особливо за складних умов, залишається непростим завданням. Комбінація стереозору з методами глибокого навчання відкриває нові перспективи для підвищення якості і точності оцінки відстані, оскільки дозволяє не лише ефективніше знаходити об’єкти, а й встановлювати їхню коректну відповідність на стереопарі. Використання згорткових нейронних мереж для аналізу ознак об’єктів значно підвищує надійність ідентифікації одного й того самого об’єкта на обох зображеннях. Це, у свою чергу, є основою для коректного розрахунку диспаратності та відповідного визначення відстані.

Метою даної курсової роботи є дослідження та розробка методики визначення відстані до об’єктів на зображенні з використанням методів глибинного навчання та аналізу стереопари. Це дослідження включає вивчення теоретичних основ комп’ютерного зору та глибинного навчання у контексті оцінки глибини, розробку алгоритмічного підходу, що поєднує виявлення об’єктів за допомогою

сучасних моделей, встановлення відповідності між об’єктами на стереозображеннях з використанням згорткової нейронної мережі та обчислення відстані на основі аналізу положення центрів відповідних об’єктів.

Розділ 1. ОГЛЯД СУЧАСНИХ ДОСЯГНЕНЬ У ГАЛУЗІ КОМП'ЮТЕРНОГО ЗОРУ

1.1. Стереозір в комп'ютерному зорі: основні поняття та застосування

Стереозір — це ключова технологія у сфері комп'ютерного зору, яка дозволяє отримувати тривимірну інформацію про сцену на основі двох або більше зображень, зроблених з різних точок спостереження. Даний підхід імітує бінокулярний зір людини та багатьох тварин, які оцінюють відстань до об'єктів за допомогою паралаксу — явища видимого зміщення об'єкта відносно фону при зміні положення спостерігача. Стереозір ґрунтується на використанні геометричних принципів триангуляції для визначення просторового розташування об'єктів.

1.1.1. Основні поняття та принцип роботи стереозору

Базовою конфігурацією стереосистеми є використання двох камер, розташованих на відомій відстані одна від одної, яка називається базовою лінією (baseline, B). Ці камери, як правило, мають паралельні оптичні осі та ідентичні внутрішні параметри (фокусна відстань, положення оптичного центру). Однак, навіть якщо оптичні осі не ідеально паралельні або камери мають різні параметри, ці відхилення можуть бути скориговані за допомогою процесу стерео калібрування. Калібрування визначає точне взаємне положення та орієнтацію камер (зовнішні параметри) та їх внутрішні параметри [1, с. 534-537].

Принцип роботи стереозору полягає у наступному: точка у тривимірному просторі проєктується на різні координати на матрицях сенсорів лівої та правої камер. Чим ближче об'єкт до камер, тим більше буде видиме зміщення його зображення між лівим та правим кадром (рис. 1.1). Це зміщення називається диспаратністю (d).

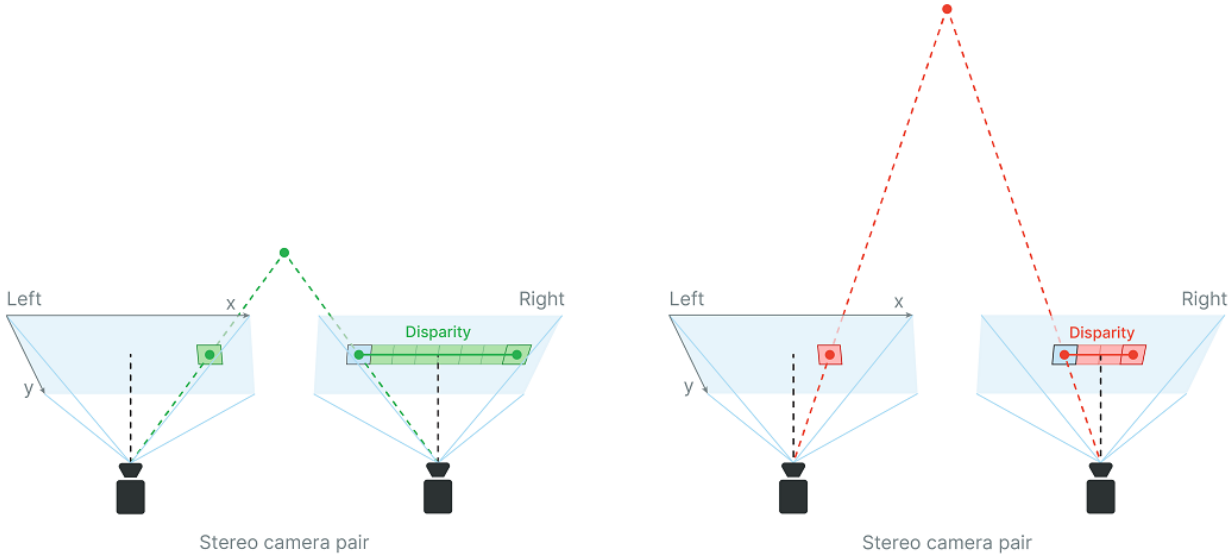


Рис. 1.1. Чим більша відстань між двома відповідними точками на лівому та правому зображенні стереопари – тим ближче об’єкт

Розглянемо ідеальний випадок паралельної стереосистеми: дві камери мають однакову фокусну відстань f, паралельні оптичні осі, що лежать на одній площині, а базові лінії сенсорів колінеарні. Нехай точка $P(X, Y, Z)$ знаходиться у 3D просторі, де Z – відстань від базової лінії камер до точки (глибина). Ця точка проєктується на координати (x_l, y_l) на лівому зображенні та (x_r, y_r) на правому зображенні. Через паралельність оптичних осей та колінеарність сенсорів, вертикальні координати проєкції будуть однаковими $(y_l = y_r)$. Горизонтальне зміщення (диспаратність) визначається як $d = x_l - x_r$.

1.1.2. Етапи класичного стереозору

Визначення тривимірної інформації про сцену за допомогою класичного пасивного стереозору складається з низки послідовних і взаємопов’язаних етапів. Перш за все, здійснюється захоплення стереопари: дві камери, розташовані на відомій відстані одна від одної, синхронно фіксують зображення тієї ж самої сцени з різних ракурсів.

Далі відбувається калібрування стереосистеми – цей етап дозволяє точно визначити внутрішні параметри кожної камери (наприклад, фокусна відстань,

координати головної точки, коефіцієнти дисторсії об’єктива) та просторові взаємодії між ними (зовнішні параметри). Результати калібрування є критично важливими: вони використовуються для корекції геометричних спотворень зображень і створюють підґрунтя для наступного етапу — ректифікації.

Ректифікація є геометричним перетворенням вихідних зображень, яке спрощує подальший пошук відповідностей шляхом вирівнювання площин зображень та оптичних осей камер таким чином, щоб проєкції будь-якої тривимірної точки на обох зображеннях лежали на одній горизонтальній лінії сканування (рис. 1.2).

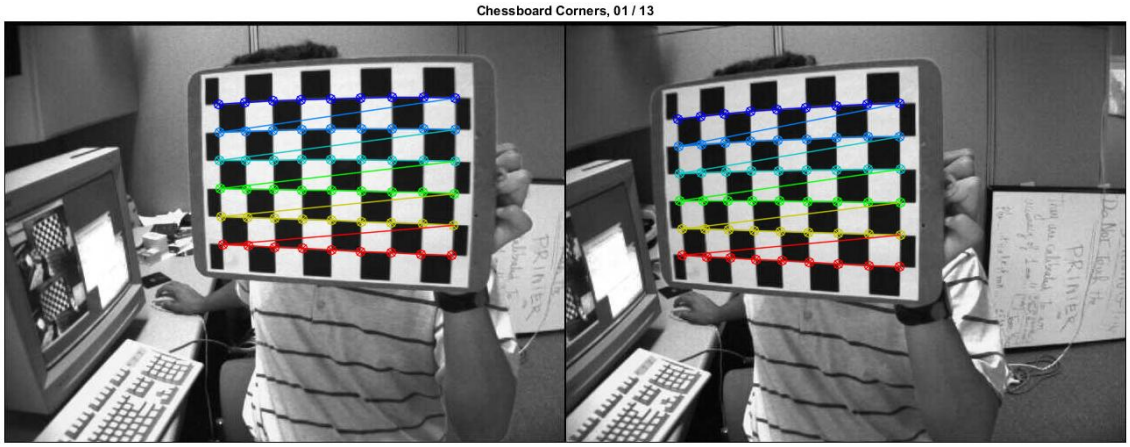


Рис. 1.2. Калібрування камер за допомогою «шахової дошки»

Після ректифікації центральним та найбільш обчислювально інтенсивним етапом є зіставлення відповідностей (correspondence matching). На цьому етапі здійснюється пошук пікселів або локальних областей (наприклад, на основі порівняння інтенсивності або характерних ознак) на лівому зображенні, які відповідають тим самим фізичним точкам у тривимірному просторі, що проєктуються на відповідні пікселі або області на правому зображенні. Точність і повнота отриманої карти глибини безпосередньо залежать від надійності та щільності знайдених відповідностей. На основі успішно знайдених відповідностей обчислюється диспаратність (d) – горизонтальне зміщення між відповідними точками на лівому та правому ректифікованих зображеннях. Результатом цього етапу є карта диспаратності, де значення кожного пікселя кодує величину цього зміщення.

Фінальним кроком є обчислення глибини (Z): отримана карта диспаратності перетворюється на карту глибини або хмару точок 3D за допомогою геометричних співвідношень триангуляції, використовуючи відомі параметри стереосистеми (базову лінію B та фокусну відстань f) за формулою $Z = \frac{B \cdot f}{d}$. Таким чином, кожен піксель на карті глибини відображає відстань до відповідної точки сцени від базової лінії камер [2].

1.2. Огляд методів визначення глибини на зображеннях

Оцінка тривимірної структури сцени на основі двовимірних зображень залишається однією з ключових та найбільш складних задач у сфері комп'ютерного зору. Визначення глибини, тобто відстані до об'єктів та окремих точок сцени, відіграє критично важливу роль у багатьох сучасних застосуваннях: автономна навігація, робототехніка, 3D-моделювання, системи безпеки, технології доповненої та віртуальної реальності. Протягом багатьох років було розроблено різноманітні підходи до вирішення цієї проблеми, і їх можна класифікувати за різними критеріями, зокрема за типом використовуваних сенсорів і за принципом отримання інформації про глибину.

Загалом, методи визначення глибини можна поділити на дві основні категорії: пасивні та активні.

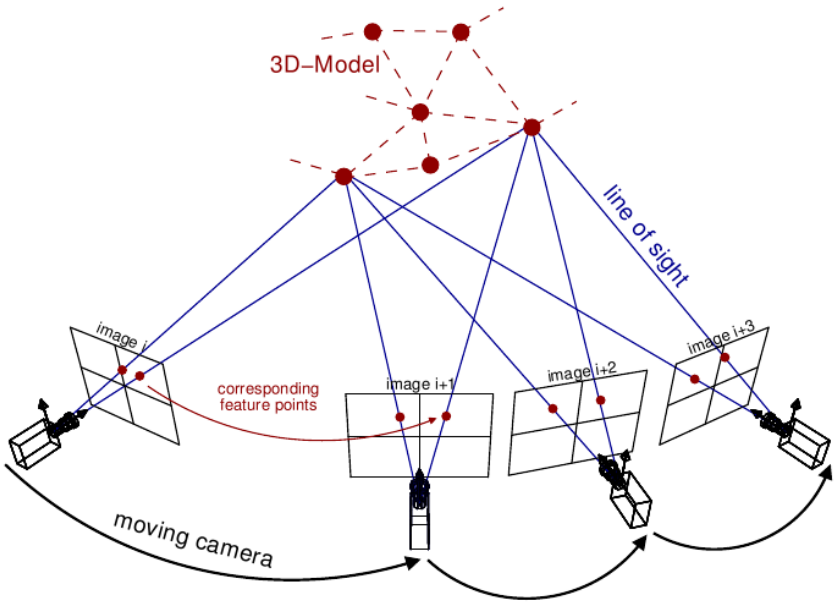
1.2.1. Пасивні методи визначення глибини

Пасивні методи використовують виключно природне або наявне освітлення сцени, не випромінюючи власну енергію. Вони покладаються на аналіз зображень, отриманих з однієї або кількох камер.

1. Стереозір: Цей метод був детально розглянутий у підрозділі 1.1.
    - Переваги: Пасивний, відносно низька вартість обладнання (дві камери), добре розроблена теорія.

- Недоліки: Складна проблема відповідності, чутливість до освітлення та відсутності текстури, проблеми з оклюзіями, обчислювальна складність.
- Приклади: Використовується у робототехніці, автономних транспортних засобах (хоча часто у поєднанні з активними датчиками), 3D-скануванні.

2. Структура з руху (Structure from Motion, SfM) та багато ракурсне стерео (Multi-View Stereo, MVS): Ці методи визначають 3D-структуру сцени та одночасно положення камери (або камер) за послідовністю зображень, отриманих з різних ракурсів. SfM фокусується на визначенні розрідженої структури сцени та траєкторії камери, тоді як MVS використовує ці дані для побудови щільної 3D-моделі або карти глибини
   - Принцип дії: Використовують алгоритми виявлення та відстеження ключових точок або ознак на зображеннях. За їх зміщенням між кадрами (паралакс) визначається відносне положення камери та 3D-координати цих точок шляхом триангуляції. MVS потім використовує епіполярну геометрію та методи зіставлення відповідностей (подібні до класичного стерео, але для багатьох ракурсів) для отримання щільної карти глибини або 3D-моделі (рис. 1.3).



2. Рис. 1.3. Графічне представлення методу SfM

- Переваги: Можливість отримання 3D-інформації з відео або набору фотографій з однієї камери (за умови руху), підходить для реконструкції великих сцен.
- Недоліки: Вимагає достатнього перекриття між зображеннями, сцени мають бути нерухомими (або рух об'єктів незначним), чутливість до текстури, обчислювально дуже інтенсивний процес, особливо MVS.
- Застосування: 3D-моделювання міст, архітектури, об'єктів для фільмів, ігор, віртуальної реальності [1, с. 347-357].

3. Монокулярне визначення глибини (Monocular Depth Estimation): Суть цього підходу полягає у спробі оцінити глибину сцени, спираючись лише на одне зображення, отримане з однієї камери. Задача апріорі не має однозначного вирішення, оскільки двовимірне зображення є проєкцією тривимірного світу, і декілька різних 3D-сцен можуть мати однакову 2D-проекцію. Саме тому монокулярні методи використовують різноманітні візуальні ознаки, що пов’язані з глибиною, наприклад, розмір

відомих об’єктів, ефекти перспективи, текстурні градієнти, особливості фокусу й дефокусу, а також затінення. Додатково застосовуються знання про типові сцени або об’єкти, які отримуються, зокрема, з навчальних даних.

- Принцип дії: Традиційні методи використовували евристики на основі класичних візуальних ознак. Сучасні підходи майже виключно базуються на глибокому навчанні, коли нейронна мережа навчається прямому або опосередкованому відображенню з 2D-зображення на карту глибини, використовуючи великі набори даних (наприклад, стереопари з відомою глибиною або дані з датчиків глибини)[3].
- Переваги: Не вимагає спеціального обладнання (крім однієї камери), підходить для пристроїв з обмеженими ресурсами.
- Недоліки: точність часто нижча, ніж у стерео або активних методів; зазвичай визначає відносну, а не абсолютну; залежить від якості та різноманітності навчальних даних.
- Застосування: Мобільні додатки з доповненою реальністю, портретний режим на смартфонах (імітація розмиття фону), деякі сценарії робототехніки.

1.2.2. Активні методи визначення глибини

Активні методи випромінюють певну форму енергії (світло, лазер, звук) у сцену та аналізують відбитий сигнал для визначення відстані.

1. Методи структурованого світла (Structured Light): Система проєктує відомий шаблон світла (наприклад, сітку, смуги або випадковий візерунок) на сцену та захоплює зображення цієї сцени однією або кількома камерами. Деформація та зміщення спроєктованого шаблону на поверхні об’єктів залежить від їх форми та відстані.

- Принцип дії: Аналізуючи деформації спроєктованого шаблону на зображенні, можна визначити просторове розташування відповідних точок за допомогою триангуляції. Такий підхід дозволяє відновити координати об’єктів у тривимірному просторі на основі їх спотворених проєкцій (рис. 1.4).

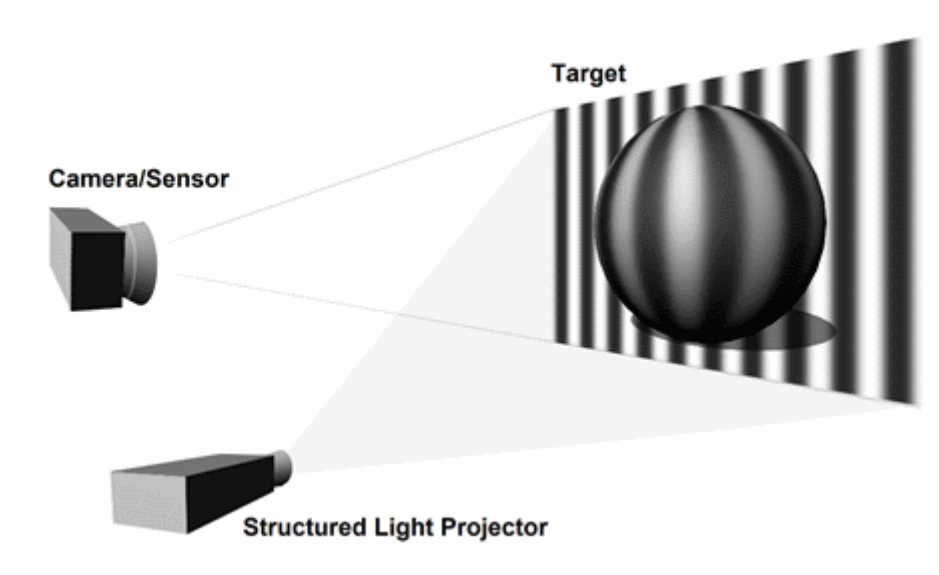


Рис. 1.4. Графічне представлення методу структурованого світла

Переваги: Висока точність на близьких відстанях, добре працює навіть на поверхнях без текстури.

- Недоліки: Чутливість до зовнішнього освітлення (особливо сонячного світла), може бути проблема з інтерференцією від інших подібних систем, зазвичай обмежений діапазон дії.
- Приклади: Microsoft Kinect (рис. 1.5), Intel RealSense (деякі моделі).

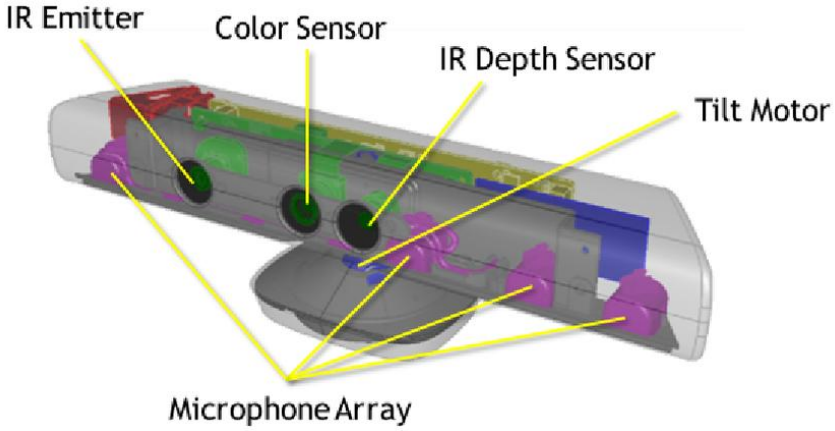


Рис. 1.5. Будова приладу Microsoft Kinect, який використовує технологію структурованого світла

2. Методи часу польоту (Time-of-Flight, ToF): Ці датчики вимірюють час, необхідний світловому імпульсу (зазвичай інфрачервоному) для досягнення об’єкта та повернення до сенсора. Знаючи швидкість світла, можна обчислити відстань.
    - Принцип дії: Датчик випромінює модульоване світло і вимірює фазовий зсув або час повернення відбитого сигналу для кожного пікселя сенсора (рис. 1.6).

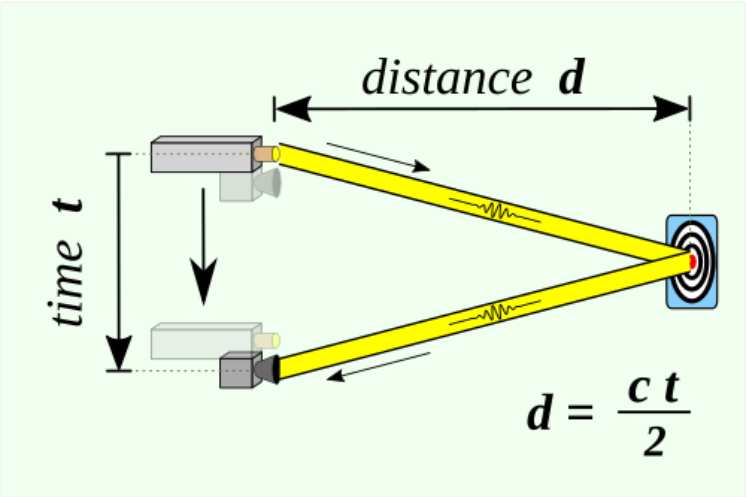


Рис. 1.6. Графічне представлення роботи методу часу польоту

    - Переваги: Пряме вимірювання глибини, висока швидкість отримання даних, менш чутливі до текстури.

- Недоліки: Зазвичай нижча просторова роздільна здатність порівняно з камерами, чутливість до розсіяного світла та ефекту багатократного відбиття, діапазон дії обмежений потужністю випромінювача.
- Приклади: Microsoft Kinect (друге покоління), промислові 3D-камери ToF [4, с. 15-20].

3. Лідар (LiDAR - Light Detection and Ranging): Лідар – це технологія, що використовує лазерні імпульси для визначення відстані до об’єктів. Суть у тому, що пристрій випромінює лазерний промінь, після чого фіксує час, за який відбитий сигнал повертається. На основі цього часу система розраховує відстань до точки, на яку був спрямований промінь. Завдяки механічному або електронному скануванню сцени формується тривимірна хмара точок, яка відображає просторову структуру об’єктів.
    - Принцип дії: Аналогічний ToF, але часто використовує скануючий промінь та більш потужні лазери для більшого діапазону.
    - Переваги: Висока точність вимірювання відстані, великий діапазон дії, добре працює в різних умовах освітлення, не залежить від текстури поверхні.
    - Недоліки: Дані, як правило, розріджені (хмара точок, а не щільна сітка), обладнання може бути дорогим та громіздким, чутливість до погодних умов (дощ, сніг, туман).
    - Застосування: Автономні транспортні засоби, високоточне 3D-картографування, геодезія, моніторинг навколишнього середовища [4, с. 5-15].

Огляд сучасних методів визначення глибини на зображеннях свідчить про значну різноманітність підходів, кожен з яких базується на власних принципах і має як переваги, так і певні обмеження. Пасивні методи аналізують лише візуальні дані, отримані при природному освітленні. Їхня гнучкість та відносна економічність на етапі впровадження роблять ці підходи привабливими для багатьох

застосувань. Водночас вони залишаються чутливими до змін умов освітлення, особливостей текстури поверхні й труднощів при пошуку відповідностей.

На відміну від пасивних, активні методи застосовують власні джерела енергії і фіксують характеристики відбитого сигналу для визначення відстані. Це дозволяє отримувати точні результати незалежно від зовнішнього освітлення чи текстурних особливостей об’єктів, проте такі системи зазвичай потребують спеціалізованого, більш вартісного обладнання. Крім того, активні підходи можуть мати обмеження щодо діапазону вимірювань або щільності отриманих даних.

1.3. Використання нейронних мереж для задач глибини та розпізнавання об’єктів

Останнє десятиріччя стало періодом глибоких змін у сфері комп’ютерного зору, передусім завдяки впровадженню глибинного навчання та сучасних нейронних мереж. Замість трудомісткого ручного проєктування дескрипторів, дослідники отримали змогу автоматично виявляти складні ієрархічні ознаки без додаткового втручання, що дозволило суттєво підвищити ефективність різноманітних завдань, зокрема розпізнавання об’єктів і визначення глибини сцени. Саме ці задачі стали ключовими напрямками розвитку, оскільки глибокі нейронні мережі, особливо згорткові (CNN), відкрили нові можливості для розуміння та аналізу візуальної інформації.

1.3.1. Нейронні мережі для розпізнавання об’єктів

Розпізнавання об’єктів передбачає ідентифікацію та локалізацію об’єктів певних класів на зображенні. У минулому для цього застосовувалися методи ручного виділення ознак, такі як SIFT та HOG, у поєднанні з класичними класифікаторами на кшталт SVM. Проте з появою глибоких згорткових нейронних мереж відбувся суттєвий прорив: сучасні підходи до розпізнавання об’єктів значно змінилися, дозволяючи автоматизувати процес виділення ознак і підвищити точність і ефективність аналізу зображень.

Згорткові нейронні мережі виявилися надзвичайно ефективними для аналізу візуальних даних завдяки своїй архітектурі, яка імітує організацію зорової кори біологічних систем. Основні компоненти CNN включають:

- Згорткові шари (Convolutional Layers): Застосовують фільтри (ядра) для виконання операції згортки із вхідним зображенням або картами ознак, вивчаючи локальні просторові ієрархії ознак – від простих країв та кутів на нижчих шарах до складніших текстур та частин об’єктів на вищих шарах (рис. 1.7).

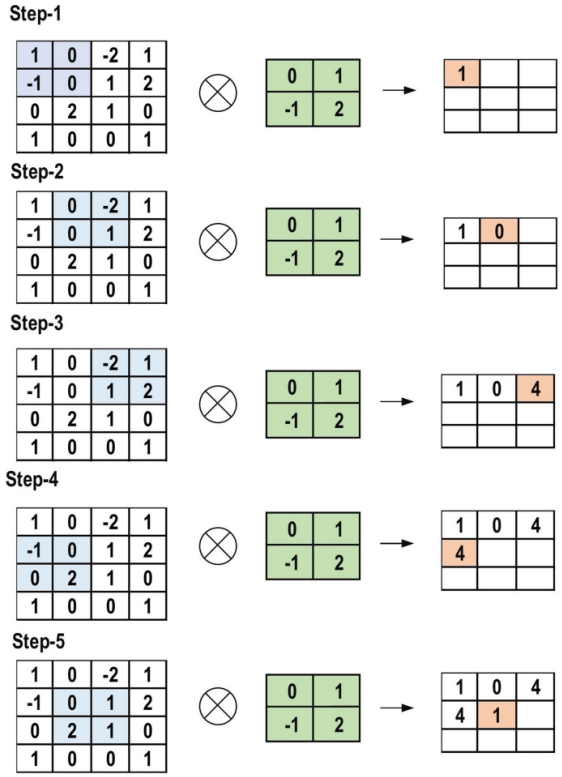


Рис. 1.7. Схематичне зображення роботи згортки

- Шари пулінгу (Pooling Layers): Виконують просторову дискретизацію (наприклад, агрегування максимальних або середніх значень у вікні), зменшуючи просторові розміри карт ознак, що допомагає зменшити обчислювальне навантаження та надає моделі інваріантності до невеликих зсувів або деформацій.

- Повнозв’язні шари (Fully Connected Layers): Знаходяться наприкінці мережі та використовують вивчені просторові ознаки для виконання кінцевої задачі – класифікації, регресії обмежувальних рамок тощо.

Нейронні мережі застосовуються для вирішення кількох пов’язаних задач розпізнавання об’єктів:

- Класифікація зображень (Image Classification): Присвоєння єдиної мітки класу всьому зображенню.
- Детекція об’єктів (Object Detection): Виявлення всіх екземплярів об’єктів певних класів на зображенні та визначення їх точного положення за допомогою обмежувальних рамок (bounding boxes), як зображено на рис. 1.8.

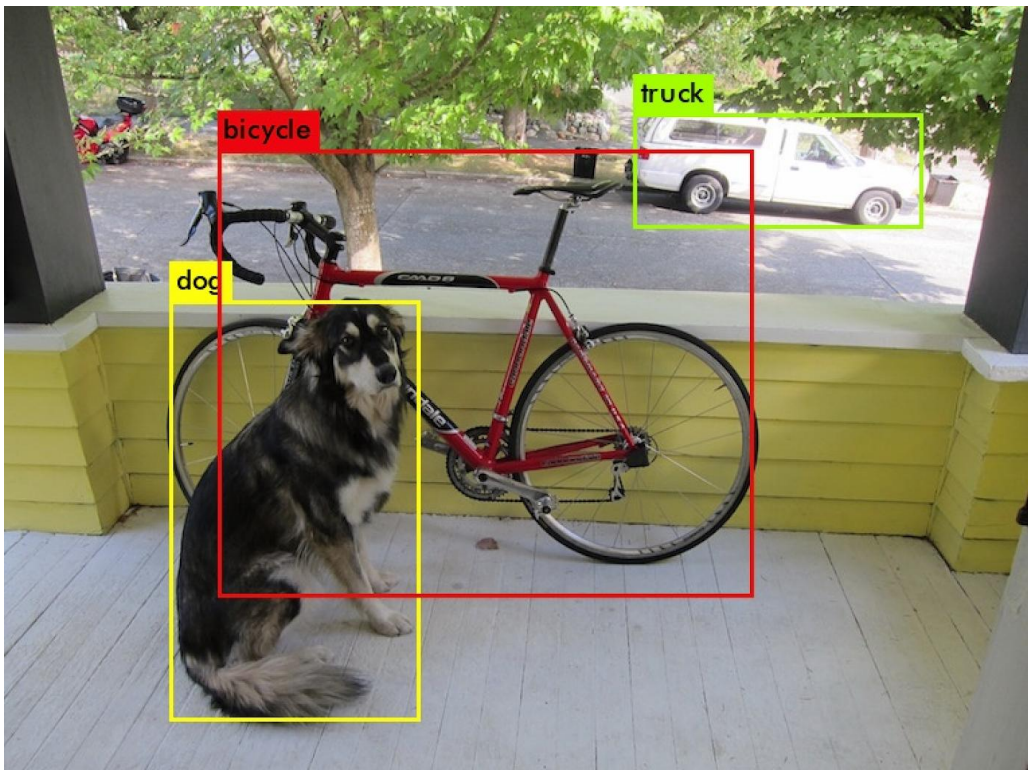


Рис. 1.8. Приклад визначення об’єктів

Сучасні методи детекції об’єктів на базі глибокого навчання поділяються на:

- Двоетапні (Two-stage detectors), які спочатку пропонують регіони, де можуть знаходитись об’єкти (наприклад, R-CNN, Fast R-CNN, Faster

R-CNN, Mask R-CNN), а потім класифікують ці регіони та уточнюють рамки.

- Одно етапні (One-stage detectors), які виконують класифікацію та регресію рамок одночасно для різних положень та масштабів на зображенні (наприклад, YOLO, SSD). Ці методи часто швидші і краще підходять для застосувань у реальному часі.

- Сегментація об’єктів (Object Segmentation): Це складніше завдання, що передбачає визначення точної піксельної маски для кожного окремого об’єкта (instance segmentation) або класифікацію кожного пікселя зображення відповідно до класу об’єкта (semantic segmentation). На відміну від простого виділення області, сегментація об’єктів дозволяє отримати більш детальне уявлення про структуру сцени.

Навчання нейронних мереж для розпізнавання об’єктів вимагає великих наборів даних із зображеннями, анотованими мітками класів та/або координатами обмежувальних рамок. Процес навчання відбувається шляхом мінімізації функції втрат (що включає втрати класифікації та втрати локалізації) за допомогою алгоритмів оптимізації, таких як градієнтний спуск.

1.3.2. Нейронні мережі для визначення глибини

У попередні роки методи визначення глибини зазвичай спиралися на явне моделювання геометрії, як-от стереозображення, або на аналіз візуальних ознак за допомогою спеціально розроблених алгоритмів, таких як SfM чи монокулярні евристики. Проте із впровадженням нейронних мереж акцент змістився у бік підходів, керованих даними. Нейронна мережа тепер має змогу самостійно навчатися оцінювати глибину чи диспаратність без необхідності ручного втручання в процес розробки ознак.

Глибоке навчання використовується для визначення глибини у кількох варіантах:

- Монокулярна оцінка глибини: Замість формалізації всіх складних монокулярних ознак глибини (перспектива, розмір, затінення) в алгоритмах, нейронна мережа навчається відображати одне 2D-зображення на карту глибини у режимі наскрізного (end-to-end) навчання. Це навчання може бути:
    - Навчання з учителем: Вимагає наборів даних із зображеннями та відповідними «істинними» картами глибини (наприклад, отриманими з LiDAR або ToF сенсорів).
    - Навчання без учителя: Використовує геометричні обмеження як сигнал помилки. Наприклад, для монокулярної оцінки можна використовувати відео послідовність, де мережа навчається оцінювати глибину та рух камери таким чином, щоб синтезовані зображення з інших ракурсів відповідали реальним кадрам.
    - Навчання з само наглядом: Часто базується на стереопарах, де ліве зображення використовується як вхід, а праве – для створення сигналу нагляду (мережі пропонується оцінити диспаратність/глибину лівого зображення, а потім «перетворити» його на праве зображення за допомогою оціненої глибини і порівняти з реальним правим зображенням).
- Стерео оцінка глибини (диспаратності): Нейронні мережі значно покращили як окремі етапи, так і всю послідовність класичного стереозору:
    - Вивчення ознак: CNN можуть навчатись витягувати більш дискримінантні ознаки для зіставлення відповідностей, ніж традиційні дескриптори.
    - Обчислення вартості відповідності: Нейронна мережа може бути навчена оцінювати «вартість» зіставлення двох фрагментів з лівого та правого зображень, що допомагає вирішити проблему відповідності.
    - Наскрізні стереомережі: Сучасні архітектури (наприклад, DispNet, PSMNet, RAFT-Stereo) приймають на вхід стереопару і безпосередньо на виході генерують щільну карту диспаратності. Ці мережі часто

включають етапи витягування ознак для обох зображень, побудову об’ємів відповідності та регресію диспаратності [5].

Розпізнавання об’єктів і визначення глибини нерозривно пов’язані між собою у контексті аналізу тривимірних сцен. Глибокі нейронні мережі демонструють універсальність, оскільки здатні виконувати обидві ці задачі або слугувати екстракторами ознак. У практиці часто застосовують архітектури, попередньо навчені на великих наборах даних для класифікації чи детекції, таких як ImageNet або COCO, які потім адаптують для вирішення задач оцінки глибини чи спеціалізованої детекції. Це дозволяє ефективно використовувати вже наявні знання мереж для нових завдань.

1.4. Огляд архітектур YOLO та їх ефективність у задачах детекції

У завданнях комп’ютерного зору, особливо під час виявлення та локалізації об’єктів на зображеннях, ключову роль відіграють нейронні мережі, здатні забезпечити обробку в реальному часі. Серед сучасних підходів до детекції об’єктів значною популярністю користуються архітектури сімейства YOLO (You Only Look Once), які вирізняються високою швидкістю роботи у поєднанні з прийнятним рівнем точності.

1.4.1. Загальна концепція YOLO

YOLO – це одностадійна архітектура глибокого навчання, яка здійснює виявлення об’єктів шляхом розгляду задачі як єдиної регресійної проблеми. На відміну від двостадійних підходів, таких як R-CNN чи Fast R-CNN, де спочатку формуються області-кандидати, а вже потім виконується класифікація, YOLO аналізує зображення за один прохід нейронної мережі. Завдяки цьому досягається значно вища швидкість роботи, що є критично важливим для застосувань у реальному часі [6].

1.4.2. Принцип роботи базової архітектури YOLO (YOLOv1)

Спершу, вхідне зображення ділиться на сітку розміром S×S. Наприклад, у YOLOv1 використовувалася сітка 7x7. Кожна комірка цієї сітки відповідає за визначення об'єктів, центри яких потрапляють у цю комірку. Для кожної комірки мережа прогнозує:

- Обмежувальні рамки (bounding boxes): Кожна рамка описується чотирма координатами $(x_{min}, y_{min}, x_{max}, y_{max})$, де (x_{min}, y_{min}) – координати верхнього лівого кута комірки сітки, а (x_{max}, y_{max}) – координати нижнього правого кута комірки сітки. Тоді центр знаходиться за формулами:

$$center\_x = \frac{\frac{x_{min} + x_{max}}{2}}{width\ of\ the\ image}$$

$$center\_y = \frac{\frac{y_{min} + y_{max}}{2}}{height\ of\ the\ image}$$

де *center_x* – центр по вісі абсцис, а *center_y* – центр по вісі ординат. Тоді для знаходження ширини та висоти рамки використовуються формули:

$$w = \frac{x_{max} - x_{min}}{width\ of\ image}$$

$$h = \frac{y_{max} - y_{min}}{height\ of\ image}$$

- Оцінку довіри для кожної рамки (confidence score): Ця оцінка відображає дві речі: по-перше, ймовірність того, що в цій рамці дійсно міститься об'єкт певного класу, і по-друге, точність самої рамки.
- Умовні ймовірності класів: Якщо комірка сітки містить об'єкт, мережа прогнозує ймовірність належності цього об'єкта до кожного з класів. Цей набір ймовірностей є спільним для всіх B рамок у цій комірці.

Далі формуються остаточні прогнози. Для кожної з рамок у кожній з S×S комірок обчислюється фінальна оцінка довіри для кожного класу: $P(class|object) \cdot confidence\ score$. Ця оцінка відображає ймовірність того, що рамка містить об'єкт певного класу і наскільки точна ця рамка.

Після отримання всіх прогнозів виконується фільтрація за рівнем довіри до класу, а потім застосовується алгоритм немаксимального пригнічення (Non-Maximum Suppression, NMS). Це дає змогу позбутися дублюючих рамок для одного й того ж об’єкта, залишаючи лише ті, що мають найвищу точність.

YOLOv1 базується на згортковій нейронній мережі, що складається з 24 згорткових шарів та 2 повнозв’язних шарів. Архітектура побудована на моделі GoogLeNet (без Inception-модулів) і містить чергування згорткових шарів, підсемплінгу (max-pooling), активаційних функцій та нормалізації (рис. 1.9).

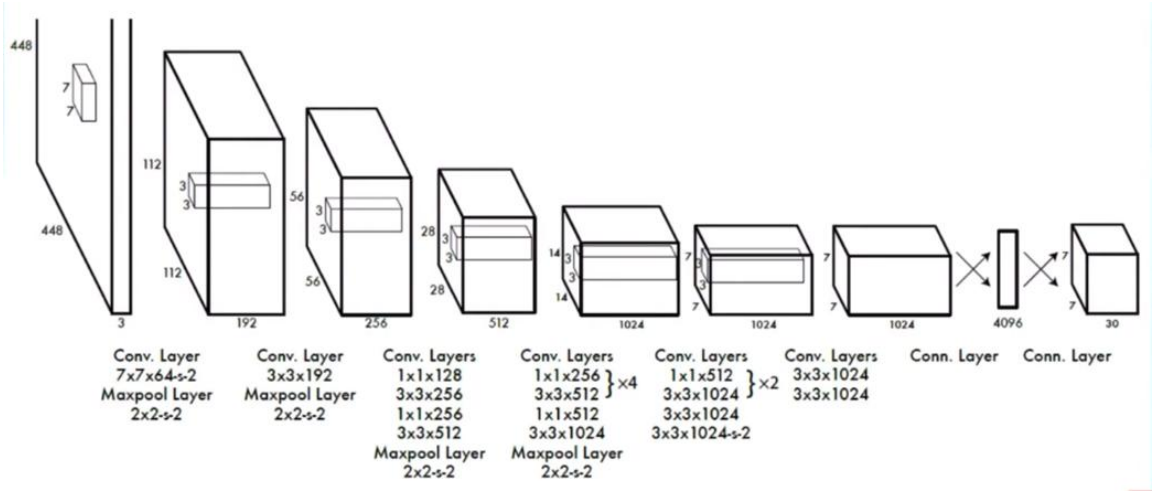


Рис. 1.9. Загальна архітектура YOLO першої версії

1.4.3. Ключові інновації та переваги YOLO

Основна особливість YOLO полягає в його одно етапному підході. Алгоритм обробляє все зображення за один прохід через нейронну мережу, на відміну від двоетапних методів. Така архітектурна відмінність забезпечує високу швидкість роботи. Варто відзначити, що YOLO став однією з перших моделей, здатних виконувати детекцію об’єктів у реальному часі на стандартному обладнанні, досягаючи понад 30 кадрів на секунду. Це, у свою чергу, відкрило нові можливості для застосування подібних систем у динамічних сценаріях.

Оскільки мережа «бачить» усе зображення під час прогнозування кожної рамки, вона краще враховує глобальний контекст сцени. Це допомагає зменшити кількість хибних спрацьовувань на фонових елементах.

1.5. Архітектура ResNet: особливості та застосування у витягуванні ознак

Історично розвиток нейронних мереж прагнув до створення все глибших архітектур, виходячи з припущення, що мережі з більшою кількістю шарів здатні вивчати складніші та абстрактніші представлення даних. Здавалося б, глибина мережі безпосередньо корелює з її можливостями. Втім, практичні спроби прямо нарощувати кількість шарів у класичних згорткових мережах, таких як VGG чи AlexNet, виявили низку серйозних викликів. Насамперед, спостерігалася проблема зникання градієнтів під час тренування, коли процес навчання фактично «зависає» через неможливість оновлення ваг. Ще одним неочікуваним явищем стала проблема деградації: при збільшенні глибини мережі спочатку фіксувалося насичення результатів на тренувальних даних, а згодом — погіршення продуктивності, і це відбувалося навіть без ознак перенавчання. Така поведінка засвідчує, що оптимізація дуже глибоких мереж є значно складнішою задачею порівняно з навчанням більш «мілких» архітектур.

Архітектура ResNet (Residual Network), представлена у 2015 році командою Microsoft Research Asia під керівництвом Каймін Хе (Kaiming He), стала ефективним рішенням проблеми деградації та дозволила успішно тренувати мережі з безпрецедентною кількістю шарів (до 152 і більше), що призвело до значного покращення результатів у багатьох задачах комп’ютерного зору.

1.5.1. Ключова інновація: Залишкові (Residual) з’єднання

Основна ідея ResNet полягає у введенні залишкових з’єднань (residual connections), також відомих як з’єднання-ярлики (shortcut connections). Замість того, щоб шари блоку намагалися безпосередньо вивчити цільове відображення $H(x)$,

де x є входом блоку, ResNet пропонує блокам вивчати залишкове відображення (residual mapping): $F(x) = H(x) - x$. Тоді вихід блоку обчислюється як сума залишкового відображення та входу:

$$H(x) = F(x) + x.$$

де x – вхідний тензор (карта ознак) блоку, $H(x)$ – бажаний вихідний тензорного блоку, а $F(x)$ – функція, що реалізується послідовністю шарів всередині блоку (зазвичай згорткові шари з активаціями та нормалізацією).

Це додавання вхідного тензора x до виходу шарів, що обчислюють $F(x)$, реалізується за допомогою залишкового з'єднання. Це з'єднання просто додає вхід блоку до його виходу перед фінальною нелінійною активацією. Якщо розмірності входу x та виходу функції $F(x)$ відрізняються, до входу x застосовується лінійне перетворення (зазвичай згортка 1x1) для приведення розмірностей перед додаванням:

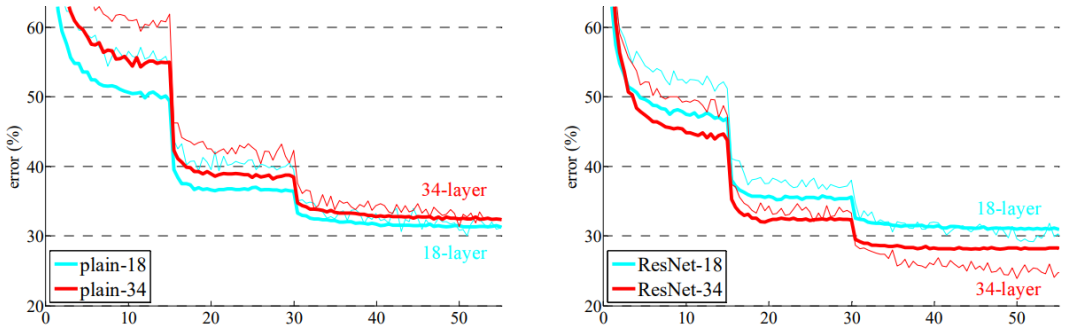
$$H(x) = F(x) + W_s x.$$

де W_s – це матриця ваг (або тензор ядер згортки 1x1) для лінійного перетворення входу x для відповідності розмірності виходу $F(x)$. У простіших випадках, коли розмірності збігаються, W_s є тотожним відображенням (множення на 1 або просто передача без змін), і формула зводиться до $H(x) = F(x) + x$.

Типовий залишковий блок ResNet складається з послідовності згорткових шарів з функціями активації (зазвичай ReLU) та пакетною нормалізацією. Якщо позначити операції згорткових шарів з нормалізацією та активацією, що обчислюють залишкову функцію $F(x)$, як F, то вихід блоку, що реалізує залишкову функцію, можна спрощено записати як $y = F(x)$. Тоді остаточний вихід залишкового блоку обчислюється шляхом додавання входу x до y та застосування фінальної нелінійної активації (наприклад, ReLU σ):

$$H(x) = \sigma(y + x) = \sigma(F(x) + x)$$

Ця проста операція додавання дозволяє ефективно «обійти» нелінійні перетворення шарів та забезпечує прямий шлях для поширення градієнтів під час зворотного поширення, значно полегшуючи тренування дуже глибоких мереж (рис. 1.10).



а *б*

Рис. 1.10. Графіки залежності похибки від кількості епох. а – прості багатошарові моделі. Модель з 34-ма шарами показує значно вищу похибку ніж модель з 18-ти шарами. б – модель ResNet. ResNet-34 з 34-ма шарами показує значно нижчу похибку ніж ResNet-18 з 18-ти шарами

1.5.2. Архітектура ResNet

Загалом, архітектура ResNet побудована так: спочатку застосовується згортковий шар разом із пулінгом, що функціонує як підготовчий етап. Далі йдуть декілька груп залишкових блоків, кожна з яких містить блоки однакового розміру карт ознак. Перехід між цими групами відбувається за допомогою згорткових шарів із певним кроком або шарів пулінгу, що дозволяє зменшувати просторові розміри та збільшувати кількість каналів ознак. Наприкінці моделі зазвичай використовується адаптивний середній пулінг, а також повнозв’язний шар, який відповідає за виконання класифікації.

Існують різні варіанти ResNet залежно від кількості шарів: ResNet-18, ResNet-34 (використовують базові блоки), ResNet-50, ResNet-101, ResNet-152 (використовують bottleneck блоки). Глибші моделі, як правило, мають більшу представницьку спроможність [7].

Розділ 2. ПРАКТИЧНА РЕАЛІЗАЦІЯ МЕТОДУ

2.1. Підготовка стереопари

Підготовка якісної стереопари є фундаментальним кроком у будь-якій системі стереозору, оскільки точність подальшого визначення глибини напряму залежить від характеристик та якості вихідних зображень. Метою цього етапу є отримання двох зображень однієї сцени з різних, чітко визначених точок огляду, що імітує бінокулярний зір людини.

2.1.1. Принцип формування стереопари

Стереозір базується на принципі триангуляції. Для визначення просторових координат точки необхідно спостерігати її щонайменше з двох різних позицій. Збільшення відстані між цими позиціями (базовою лінією) підвищує точність вимірювання глибини для віддалених об’єктів. Водночас надто велика базова лінія може ускладнити пошук відповідностей для близько розташованих об’єктів через значні зміщення.

2.1.2. Методологія підготовки стереопари

Даний підхід до створення стереопари є прикладом ручної послідовної зйомки, де одна камера використовується для захоплення двох зображень з різних позицій. Така методологія включає наступні етапи:

1. **Підготовка сцени та перше зображення (Ліве):** Спочатку розміщуються об’єкти таким чином, щоб вони утворювали тестову сцену. Смартфон OnePlus 8 з модулем камери IMX586, встановлюється перед сценою так, щоб усі об’єкти були в кадрі, і захоплюється перше фото. Це зображення вважається «лівим» у стереопарі.

2. **Зміщення камери та друге зображення (Праве):** Після цього камера зміщується вправо на n сантиметрів, де n може бути різним дійсним значенням. Ця величина n є базовою лінією, і її вибір впливатиме на точність

обчислення глибини. Критично важливим аспектом є збереження кута огляду незмінним. Це мінімізує спотворення, спричинені обертанням камери. Після зміщення робиться друге фото, яке слугує «правим» зображенням стереопари.

2.1.3. Ключові аспекти та виклики при підготовці стереопари

Незважаючи на простоту описаного підходу, його ефективність та точність подальшого обчислення глибини суттєво залежать від врахування наступних факторів:

1. Калібрування камери: Це важливий етап, оскільки більшість камер мають певні оптичні спотворення. Якщо цей процес пропустити, прямі лінії на зображенні можуть виглядати викривленими, що негативно впливає на точність визначення глибини. У цій роботі калібрування камери не виконувалося, тому можливі похибки у вимірюванні відстані.
2. Зміщення камери: Ручне виконання паралельного зміщення без спеціального обладнання (наприклад, оптичної рейки) рідко дозволяє досягти досконалості. Навіть за умови максимально обережного ручного зміщення, незначні розбіжності в орієнтації камер неминучі, що може негативно впливати на точність визначення відстані.
3. Статичність сцени: Усі об’єкти в сцені повинні бути абсолютно нерухомими між моментом зйомки першого та другого фото. Будь-який рух об’єктів призведе до невірних значень і, як наслідок, до неправильного обчислення глибини.

В кінці цього етапу отримуємо дві цифрові фотографії – ліве та праве зображення стереопари. Хоча ручний спосіб отримання стереопари не може забезпечити таку точність і відтворюваність, як спеціалізовані стерео камери чи системи з лазерними далекомірами, у межах курсової роботи цей підхід цілком прийнятний для демонстрації основних ідей і перевірки алгоритмів. Зібрані стереопари дають можливість засвоїти базові принципи роботи системи. Надалі саме ці

зображення будуть використані як вхідні дані для етапів детекції об’єктів за допомогою YOLO, співставлення ознак через ResNet і розрахунку відстані.

2.2. Розпізнавання об’єктів за допомогою моделі YOLO

Для виконання цієї задачі обрано ефективну архітектуру YOLOv8 від компанії Ultralytics. Вибір цієї моделі обумовлений її високою точністю, швидкістю та зручністю використання, що робить її ідеальним інструментом для інтеграції у проєкт визначення відстані до об’єктів.

2.2.1. Огляд Ultralytics YOLO та вибір моделі

Компанія Ultralytics активно розвиває сімейство моделей YOLO, починаючи з YOLOv3, і пропонують оптимізовані версії, які легко інтегруються та масштабуються для різних застосувань. Ultralytics YOLOv8 є однією з найпоширеніших та порівняно сучасних ітерацій, що пропонує покращену архітектуру, нові функції втрат та оптимізовані стратегії тренування, що призводить до вищої продуктивності порівняно з попередніми версіями.

Ultralytics надає різні варіанти моделей YOLOv8, які відрізняються розміром, складністю, швидкістю та точністю (рис. 2.1). Ці варіанти позначаються суфіксами:

- yolov8n (nano): Найменша та найшвидша модель. Призначена для застосувань, де критична швидкість та обмежені обчислювальні ресурси (наприклад, мобільні пристрої, вбудовані системи). Має найменшу кількість параметрів.
- yolov8s (small): Трохи більша за n версію, пропонує кращий баланс між швидкістю та точністю.
- yolov8m (medium): Збільшена модель, що забезпечує вищу точність за рахунок помірного збільшення обчислювальних витрат.

- yolov8l (large): Велика модель з високою точністю, але вимагає значних обчислювальних ресурсів.
- yolov8x (extra large): Найбільша та найточніша модель у сімействі, призначена для завдань, де точність є пріоритетом, а обчислювальні ресурси не є обмеженням.

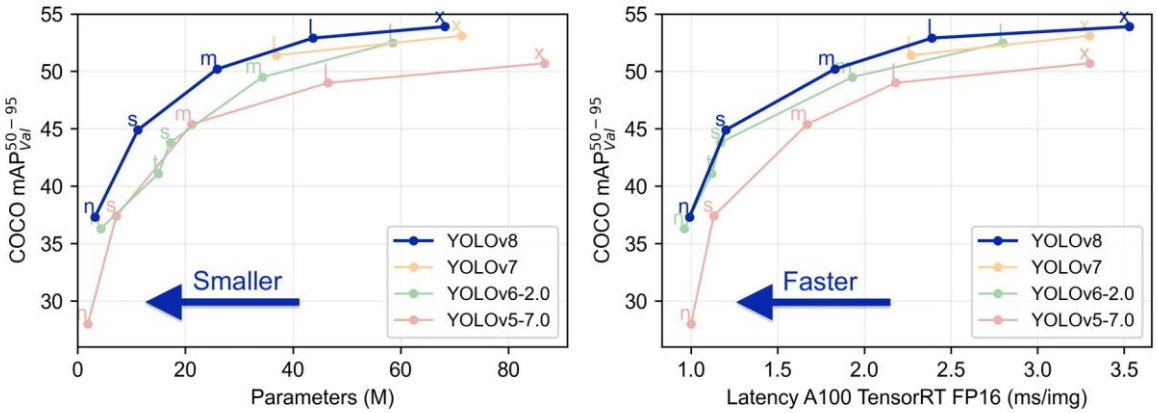


Рис. 2.1. Відношення точності до розміру та швидкості обробки зображень

Для даної курсової роботи було обрано модель yolov8n. Версія nano є найшвидшою, що дозволяє оперативно обробляти зображення стереопари та демонструвати принципи роботи в реальному часі (або наближено до нього) без потреби у потужному графічному процесорі. Незважаючи на свій невеликий розмір, yolov8n забезпечує достатню точність детекції для більшості типових об’єктів, що є адекватним для демонстраційних цілей курсової роботи. Модель yolov8n вимагає мінімальних обчислювальних ресурсів та пам’яті, що спрощує її запуск та тестування на стандартних комп’ютерах. А бібліотека ultralytics надає простий та інтуїтивно зрозумілий API для завантаження та використання моделей YOLO, що прискорює розробку.

Результатом виконання даного етапу для кожного зображення стереопари є список виявлених об’єктів. Кожен елемент цього списку містить шлях до обрізаного зображення об’єкта, його клас та точні координати обмежувальної рамки на оригінальному зображенні.

Таким чином, інтеграція YOLOv8 у систему забезпечує швидкий, точний та автоматизований спосіб підготовки даних для подальшого визначення глибини, демонструючи ефективне поєднання сучасних методів детекції об’єктів з принципами стереозору.

2.3. Витягування ознак за допомогою ResNet-18

Для витягування ознак застосовується архітектура ResNet-18. Як уже зазначалося в підрозділі 1.5, ResNet – це архітектура згорткової нейронної мережі, яка завдяки залишковим з’єднанням дозволяє ефективно тренувати дуже глибокі моделі. ResNet-18 є порівняно компактною версією цієї архітектури, що складається з 18 шарів. Незважаючи на відносно невелику кількість шарів, ця модель залишається ефективним екстрактором ознак.

ResNet-18 забезпечує хороший баланс між обчислювальною складністю та якістю витягнутих ознак. Вона достатньо мала, щоб швидко виконувати прямий прохід, що важливо для практичних застосувань, але водночас достатньо глибока, щоб вивчати багаті та абстрактні візуальні представлення. Шари ResNet-18 навчаються ієрархічним представленням: від простих країв та текстур на ранніх шарах до більш складних, семантичних ознак об’єктів на глибших шарах. Ці високорівневі ознаки є набагато більш стійкими до змін у освітленні, масштабі чи ракурсі, ніж сирі піксельні значення або традиційні дескриптори. Як і у випадку з YOLO, для ResNet-18 доступні попередньо навчені ваги на великих наборах даних (наприклад, ImageNet). Використання таких моделей дозволяє застосовувати передавальне навчання. Саме тому ResNet-18 є хорошим вибором для даної задачі.

На практичному етапі витягнення ознак та порівняння об’єктів усе відбувається наступним чином. Для початку використовується нейронна мережа ResNet-18, яка попередньо навчена на великому наборі даних (наприклад, ImageNet). Оскільки ціль полягає не у класифікації зображень, а у отриманні глибоких візуальних представлень, фінальний шар моделі, що відповідає за

класифікацію, видаляється. Внаслідок цього ResNet-18 виступає вже не як класифікатор, а як екстрактор ознак. На виході мережа формує вектор ознак – компактне числове представлення ключових візуальних характеристик вхідного зображення. Це дозволяє порівнювати різні об’єкти на основі їхніх візуальних ознак.

Перед подачею зображення об’єкта в модель, воно проходить етап стандартизації. Це включає зміну його розміру до фіксованих параметрів (наприклад, 224x224 пікселів), перетворення його у числовий тензор та нормалізацію пікселльних значень. Ці кроки є критично важливими, оскільки модель була навчена саме на таких підготовлених даних.

На наступному етапі кожне обрізане зображення об’єкта, отримане після детекції за допомогою YOLO, подається на вхід попередньо навченої моделі ResNet-18. Під час інференсу модель формує унікальний, високо розмірний вектор ознак. Об’єкти з подібними характеристиками матимуть схожі вектори ознак. Варто зауважити, що на цьому етапі ваги моделі не змінюються, отже, процес є швидким та ефективним.

Ключовим моментом є порівняння цих отриманих векторів ознак. Для кожної потенційної пари об’єктів (наприклад, один об’єкт з лівого зображення стереопари та один з правого), обчислюється міра їхньої схожості. Для цього використовується косинусна схожість, яка визначає «кут» між векторами ознак. Значення косинусної схожості, близькі до 1, вказують на високу схожість об’єктів.

На основі цього показника схожості застосовується порогова фільтрація: якщо косинусна схожість між двома об’єктами перевищує заздалегідь визначений поріг (наприклад, 0.8), вони вважаються відповідними (тобто, одним і тим ж фізичним об’єктом у просторі, видимим з двох камер). Пари, що не відповідають цьому критерію, відкидаються як хибні зіставлення. У результаті цього етапу формується список лише тих пар об’єктів, які були надійно підтверджені як відповідності, що є основним вихідним даним для подальших геометричних розрахунків глибини.

2.4. Обчислення глибини об’єкта на основі положення центрів

Після успішного розпізнавання об’єктів за допомогою YOLO та їх надійного порівняння між лівим та правим зображеннями стереопари за допомогою ResNet-18, ми маємо всю необхідну інформацію для застосування принципів стерео геометрії та розрахунку глибини. Основний принцип полягає у використанні диспаратності – видимого зміщення об’єкта між двома зображеннями – для обчислення його відстані від камер.

2.4.1. Ключові параметри для обчислення глибини

Для точного розрахунку глибини необхідні наступні параметри, які характеризують камеру та її розташування:

- Ширина зображення в пікселях (image_width): Це горизонтальна роздільна здатність зображення, отримана з модуля IMX586. Вона використовується для перетворення кутових вимірів у піксельні.
- Горизонтальний кут огляду (h_fov_deg): Це кут, який охоплює камера (IMX586) по горизонталі. Знаючи його, можна обчислити фокусну відстань камери в пікселях.
- Базова лінія в сантиметрах: Це точна відстань між оптичними центрами лівої та правої позицій камери, з яких були зроблені знімки. Цей параметр є фундаментальним для триангуляції.

2.4.2. Принцип роботи обчислення глибини

Процес обчислення глибини для кожної підтвердженої пари об’єктів відбувається за наступними кроками:

1. Обчислення фокусної відстані в пікселях: Перед початком розрахунку глибини необхідно визначити фокусну відстань камери в пікселях (*f*), оскільки цей параметр є ключовою внутрішньою характеристикою камери. Він визначає, як розміри об’єктів у реальному світі співвідносяться з їх

відображенням у пікселях на зображенні. Для обчислення фокусної відстані використовують ширину зображення та горизонтальний кут огляду камери. Цей розрахунок проводиться лише один раз, оскільки параметри камери залишаються постійними для всіх об'єктів. При цьому важливо враховувати, що горизонтальний кут огляду слід виражати у радіанах.

$$h\_fov\_rad = \frac{h\_fov\_deg \cdot \pi}{180}$$

де h_fov_deg – горизонтальний кут огляду камери. Після цього можна обчислювати фокусну відстань в пікселях (*f*):

$$f = \frac{image\_width}{2 \cdot tan(\frac{h\_fov\_rad}{2})}$$

2. Визначення центрів об'єктів та диспаратності: Для кожної пари об'єктів, які були успішно зіставлені на попередньому етапі (за допомогою ResNet-18), ми маємо їхні обмежувальні рамки на лівому та правому зображеннях. Щоб обчислити диспаратність, необхідно визначити горизонтальну координату центральної точки кожного об'єкта в пікселях. Центр об'єкта обчислюється як середина його обмежувальної рамки по горизонталі. Диспаратність (*d*) потім розраховується як абсолютна різниця між горизонтальною координатою центру об'єкта на лівому зображенні та горизонтальною координатою центру того ж об'єкта на правому зображенні. Важливо, що для коректного розрахунку диспаратність завжди є позитивним числом. Чим ближче об'єкт, тим більшою буде диспаратність, оскільки його видиме зміщення між зображеннями буде значнішим.
3. Застосування формули триангуляції: Після обчислення фокусної відстані та диспаратності, глибина об'єкта (Z) розраховується за фундаментальною формулою триангуляції, яка була розглянута в підрозділі 1.1.2.
4. Обробка особливих випадків: Існує важливий особливий випадок: якщо диспаратність дорівнює нулю. Це може статися, якщо об'єкт знаходиться

на дуже великій відстані (теоретично, на нескінченності), або якщо виникла помилка у зіставленні відповідностей. У такому випадку, ділення на нуль неможливе, і відстань не може бути обчислена.

2.4.3. Роль у курсовій роботі

Цей етап є кульмінацією всієї роботи. Він перетворює візуальні дані та виявлені відповідності на кількісну інформацію про тривимірне розташування об’єктів. Точність визначення глибини безпосередньо залежить від якості попередніх етапів: точності отриманої стереопари, коректності розпізнавання об’єктів за допомогою YOLO, а також достовірності їхньої верифікації через ResNet-18. Основний результат цього етапу – це обчислені відстані до кожного з виявлених об’єктів, які й становлять ключове вихідне значення в усьому процесі.

2.5. Тестування підходу

Для тестування було підготовлено декілька стереопар зображень, захоплених згідно з методологією, описаною в підрозділі 2.2. Кожна стереопара складається з лівого та правого зображень однієї й тієї ж сцени, отриманих зі зміщенням камери на відому базову лінію.

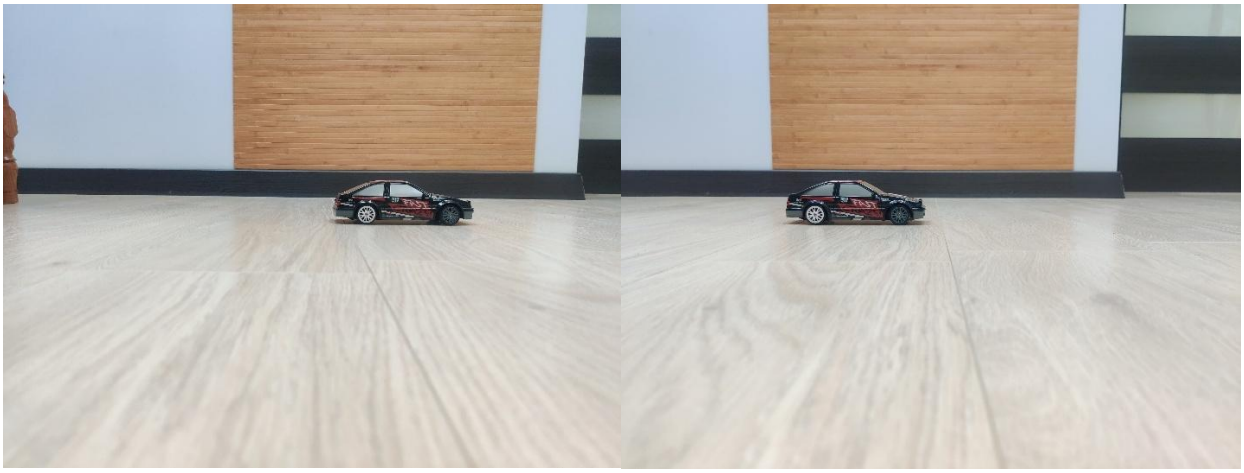
Приклад 1: Сцена з одним об’єктом

Опис сцени: У цій сцені розташовано один об’єкт (іграшкова модель авто). Це дозволить перевірити здатність системи оцінювати глибину для об’єкта. Тестування відбувалось на двох відстанях до об’єкта: 30 см. та 60 см.

1. Вхідні зображення на рис. 2.5.1 та 2.5.2.

а) б)

Рис. 2.5.1. Вхідні зображення на відстані 30 см.: а) ліве зображення стереопари, б) праве зображення стереопари



а) б)

Рис. 2.5.1. Вхідні зображення на відстані 60 см.: а) ліве зображення стереопари, б) праве зображення стереопари

2. Результати зображені на рис. 2.5.3 та 2.5.4.

Рис. 2.5.3. Результат обчислення програми для зображення, де об’єкт знаходився на відстані 30 см.

Рис. 2.5.4. Результат обчислення програми для зображення, де об’єкт знаходився на відстані 60 см.

3. Аналіз результатів (табл. 2.5.1)

Табл. 2.5.1

Очікувана відстань (см)	**Практична відстань (см)**	**Абсолютна похибка (см)**	**Відносна похибка (%)**
30.0	30.8	0.8	2.67
60.0	59.3	0.7	1.17

У проведених випробуваннях отримані значення відстаней продемонстрували високу точність, що підтверджується низькими абсолютними та відносними похибками.

Зокрема, при очікуваній відстані 30 см система визначила 30,8 см. Абсолютна похибка склала 0,8 см, а відносна — 2,67%. Це свідчить про високу точність вимірювання на малих відстанях, де більша диспаратність зменшує вплив незначних помилок у піксельних вимірах.

Щодо об’єкта, розташованого на відстані 60 см, практичне значення дорівнювало 59,3 см. Абсолютна похибка становила 0,7 см, а відносна — лише 1,17%. Варто зазначити, що навіть при збільшенні відстані у два рази абсолютна похибка залишилася практично на тому ж рівні, а відносна навіть зменшилася. Це підтверджує стабільність і надійність системи на різних відстанях.

Приклад 2: Сцена з кількома об’єктами на різних відстанях

Опис сцени: У цій сцені розташовано кілька об’єктів різного типу на різних відстанях від камери. Це дозволить перевірити здатність системи визначати різноманітні об’єкти та оцінювати глибину для об’єктів, що знаходяться на різних планах.

1. Вхідні зображення на рис. 2.5.5.

а) б)

Рис. 2.5.5. Вхідні зображення: а) ліве зображення стереопари, б) праве зображення стереопари

2. Результати зображені на рис. 2.5.6.

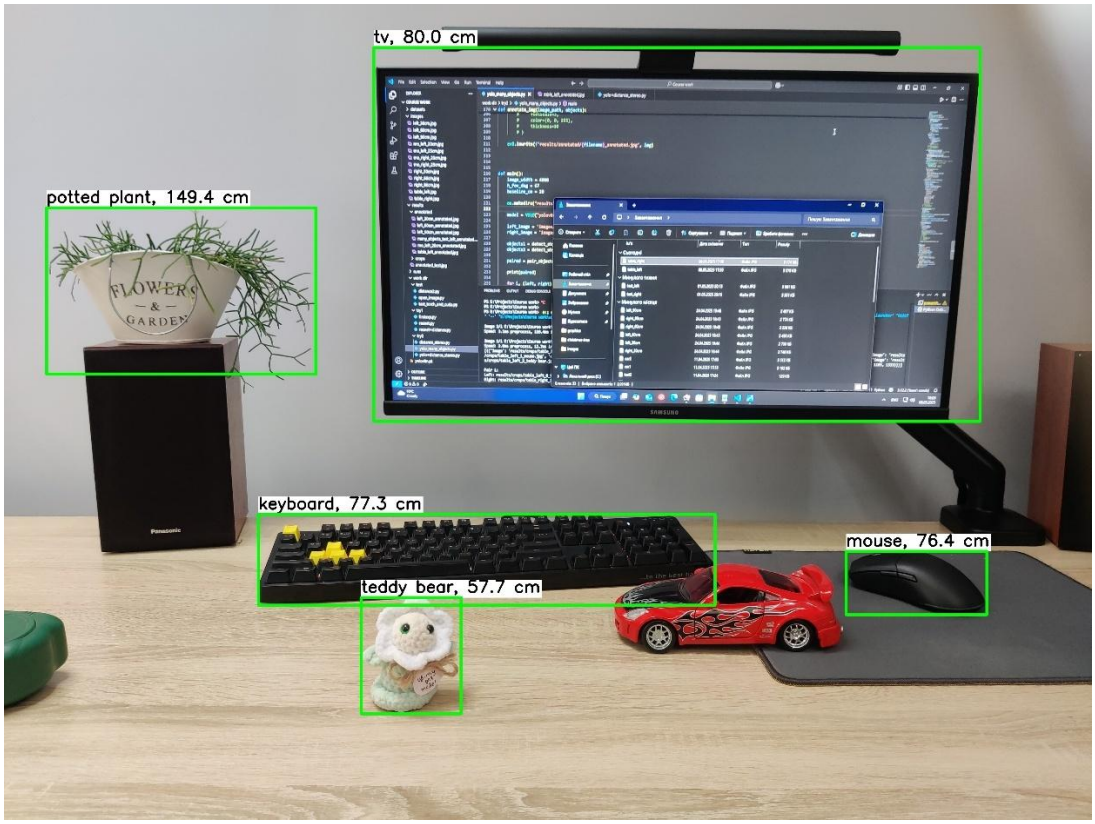


Рис. 2.5.6. Результат обчислення програми для зображення з багатьма об’єктами.

3. Аналіз результатів (табл. 2.5.2)

Табл. 2.5.2

Об’єкт	**Очікувана відстань (см)**	**Практична відстань (см)**	**Абсолютна похибка (см)**	**Відносна похибка (%)**
В’язана іграшка	~61.0	57.7	3.3	5.41
Комп’ютерна миша	~80.0	76.4	3.6	4.5
Клавіатура	~84.0	77.3	6.7	7.98
Монітор	~91.0	80.0	11.0	12.09
Іграшкова модель авто	~69.0	Об’єкт не виявлено	N/A	N/A
Кімнатна рослина	~80-90	149.4	~60-70	~40-50

Представлені результати тестування дозволяють зробити кілька важливих висновків щодо ефективності та обмежень розробленої методики визначення відстані до об’єктів.

1. Загальна точність: Для більшості виявлених і коректно ідентифікованих об’єктів, таких як в’язані іграшки, комп’ютерна миша та клавіатура, система показала достатньо прийнятну точність — відносна похибка коливалася у межах від 4,5% до 8%. Це можна вважати хорошим результатом для системи, яка ґрунтується на ручному налаштуванні стереопари та застосуванні попередньо навчених нейронних мереж.
2. Вплив повноти об’єкта в кадрі: Випадок з кімнатною рослиною чітко демонструє критичну залежність точності від повноти видимості

об’єкта на обох зображеннях. Якщо на одному з кадрів об’єкт значно обрізаний (див. рис. 2.5.7), визначення центру його обмежувальної рамки стає некоректним. Це, у свою чергу, призводить до суттєвих помилок у розрахунку диспаратності, а отже – й до повністю хибного визначення відстані.

а) б)

Рис. 2.5.7.Зображення кімнатної рослини зі стереопари: а) ліве зображення стереопари, б) праве (обрізане) зображення стереопари

3. Проблеми детекції: Випадок з іграшковою моделлю авто, яка не була виявлена, вказує на обмеження самого етапу детекції об’єктів YOLO. Це може бути пов’язано з моделюю версії «nano», яка є найменшою та найпростішою з моделей.

Висновок

У рамках цієї курсової роботи було розроблено комплексний підхід до визначення відстані до об’єктів, що поєднує методи стереозору та глибокого навчання. Основна мета полягала у поєднанні сучасних нейронних мереж для обробки зображень із класичними принципами стерео геометрії, що дало змогу досягти автоматизованих і надійних результатів.

Було розглянуто процес ручного захоплення стереопари за допомогою смартфона, з акцентом на забезпеченні паралельного зміщення. Цей етап заклав основу для подальших розрахунків, підкреслюючи важливість точності базової лінії.

Для автоматичної ідентифікації та локалізації об’єктів на обох зображеннях стереопари використовувалася легка та швидка модель YOLOv8n. Цей етап забезпечив обмежувальні рамки та класові мітки для кожного виявленого об’єкта, що є критично важливим для наступних кроків.

Для надійного зіставлення об’єктів між лівим і правим зображеннями були використані глибокі семантичні ознаки, отримані за допомогою моделі ResNet-18. Застосування косинусної схожості дало змогу підтвердити відповідності, ефективно відфільтровуючи помилкові спрацювання.

На основі зіставлених центрів об’єктів, відомої базової лінії та параметрів камери, була обчислена диспаратність. За допомогою формули триангуляції ця диспаратність була перетворена на метричну відстань до кожного об’єкта.

Проведене тестування на зображеннях підтвердило працездатність та ефективність розробленого підходу. Для більшості об’єктів, що були коректно виявлені та повністю видимі, система продемонструвала високу або прийнятну точність визначення відстані (з відносною похибкою в діапазоні 1-8%).

Однак, робота також виявила певні обмеження та виклики. Якщо об’єкти частково обрізані на зображенні або модель YOLO їх не розпізнає, обчислення

відстані стає некоректним або взагалі неможливим. Це чітко демонструє важливість ретельної підготовки стереопари й надійної роботи алгоритму детекції. Водночас, ручне захоплення стерео знімків є зручним для демонстрації, проте не гарантує ідеальних умов, зокрема паралельності камер чи точності вимірювання базової лінії, що в підсумку впливає на точність результатів. Крім того, використання спрощених моделей YOLO (наприклад, версії v8n) для підвищення швидкості іноді позначається на їх здатності виявляти дуже малі або незвичайні об’єкти.

Незважаючи на зазначені обмеження, дана курсова робота закладає основу для подальших досліджень. Можливі напрямки вдосконалення включають:

- Автоматичне калібрування камери: Впровадження алгоритмів автоматичного калібрування камер для компенсації недоліків викривлення зображення.
- Використання більших моделей: Тестування з більшими версіями YOLO (наприклад, YOLOv8m) для підвищення точності детекції.
- Інтеграція з реальними стерео камерами: Перехід до використання професійних стерео камер для отримання більш точних та синхронізованих стереопар.

У підсумку, дана курсова робота чітко ілюструє, як класичні підходи комп’ютерного зору можуть ефективно поєднуватися з сучасними методами глибинного навчання для вирішення практичних задач, зокрема визначення відстані до об’єктів. Отримані результати відкривають перспективи для подальшої розробки більш просунутих та автономних візуальних систем.

Джерела

1. Computer Vision: Algorithms and Applications. Springer, 2010. 832 с.
2. Multiple view geometry in computer vision. Cambridge University Press, 2000. 624 с.
3. Prince S. J. D. Computer vision: models, learning, and inference. Cambridge University Press, 2012. 665 с.
4. An overview of depth cameras and range scanners based on time-of-flight technologies / R. Horaud та ін. *Machine vision and applications*. 2016. Т. 27, № 7. С. 1005–1020. URL: https://doi.org/10.1007/s00138-016-0784-4 (дата звернення: 19.05.2025).
5. Forsyth D. A., Ponce J. Computer vision: a modern approach. Pearson Education, Limited, 2015. 793 с.
6. Object detection using yolo algorithm. *International research journal of modernization in engineering technology and science*. 2024. URL: https://doi.org/10.56726/irjmets60812 (дата звернення: 19.05.2025).
7. Rosebrock A. Deep learning for computer vision with python. PYIMAGESEARCH, 2017. 321 с. URL: https://bayanbox.ir/view/5130918188419813120/Adrian-Rosebrock-Deep-Learning-for.pdf (дата звернення: 19.05.2025).