МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДОНЕЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ ВАСИЛЯ СТУСА
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ І ПРИКЛАДНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КОСТЮШИН МАКСИМ ВІКТОРОВИЧ

Допускається до захисту:
в.о. завідувача кафедри
інформаційних технологій,
д-р. техн. наук, професор
_____Наталія ВЕСЕЛОВСЬКА
«____»_____ 2025 р.

МУЛЬТИМОДАЛЬНЕ РОЗПІЗНАВАННЯ ОБ'ЄКТІВ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

Спеціальність 122 Комп'ютерні науки

Кваліфікаційна (магістерська) робота

Науковий керівник:
Петро НІКОЛЮК, професор кафедри
інформаційних технологій,
д-р фіз.-мат. наук, професор
_____

Оцінка: _____ / _____ / _____
(бали/за шкалою ЄКТС/за національною шкалою)

Голова ЕК: _____
(підпис)

Вінниця – 2025

АНОТАЦІЯ

Костюшин М.В Мультимодальне розпізнавання об’єктів з використанням штучного інтелекту. Спеціальність 122 «Комп’ютерні науки», освітня програма «Комп’ютерні технології обробки даних» . Донецький національний університет імені Василя Стуса, Вінниця 2025.

Магістерська робота присвячена дослідженню предметної області та практичній реалізації мультимодальної системи розпізнавання об’єктів на основі поєднання RGB-зображень та карти глибини. У роботі проведено аналіз сучасного стану мультимодального навчання, розглянуто ключові архітектури, алгоритми та теоретичні основи представлення, злиття та вирівнювання модальностей. На основі цього сформульовано технічні вимоги до системи та обґрунтовано вибір методів і інструментів. Робота формує основу для подальшого вдосконалення технік мультимодального злиття та розширення системи іншими типами даних.

Ключові слова: мультимодальність, штучний інтелект, розпізнавання об’єктів

92 с., 1 табл., 20 рис., 1 дод., 56 джерел

ABSTRACT

Kostyushyn M.V. Multimodal object recognition using artificial intelligence. Specialty 122 "Computer Science", educational program "Computer Data Processing Technologies". Vasyl Stus Donetsk National University, Vinnytsia 2025.

The master's thesis is devoted to the study of the subject area and practical implementation of a multimodal object recognition system based on a combination of RGB images and a depth map. The paper analyzes the current state of multimodal learning, considers key architectures, algorithms and theoretical foundations of representation, fusion and alignment of modalities. Based on this, technical requirements for the system are formulated and the choice of methods and tools is justified. The work forms the basis for further improvement of multimodal fusion techniques and expansion of the system with other data types.

Keywords: multimodality, artificial intelligence, object recognition

92 p., 1 table, 20 figures, 1 appendix, 56 sources

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ

RGB – Формат подання зображення за трьома каналами кольору

D – Представлення сцени у вигляді карти глибини

RGB+D – Комбіноване подання кольорового зображення та карти глибини

CNN – Тип нейронної мережі, що використовується для обробки зображень

YOLO – Модель для швидкої детекції об’єктів у зображеннях

MiDaS – Модель для оцінки відносної глибини зі звичайного зображення

mAP – Метрика якості детекції, що оцінює середню точність за всіма класами

ВСТУП

У сучасному світі, де обсяги інформації зростають з кожним днем, ключову роль відіграють системи, здатні розуміти, аналізувати та інтерпретувати дані різних типів. Одним із найдинамічніших напрямів розвитку штучного інтелекту є мультимодальне розпізнавання об’єктів — підхід, що поєднує в собі різні джерела інформації. Людське сприйняття є мультимодальним за своєю природою — ми одночасно використовуємо зорові, слухові, текстові та тактильні сигнали для розпізнавання об’єктів і прийняття рішень. Аналогічно, мультимодальні моделі штучного інтелекту дозволяють комп’ютерам комбінувати дані з різних сенсорних каналів (зображення, аудіо, текст, відео) для підвищення точності розпізнавання та розуміння контексту.

Сучасний розвиток штучного інтелекту спрямований на створення систем, здатних комплексно сприймати, аналізувати та інтерпретувати інформацію з різних джерел. Завдяки швидкому прогресу в галузях комп’ютерного зору, глибокого навчання та мовних моделей, мультимодальні системи поступово стають фундаментом для нових поколінь інтелектуальних технологій — від автономних роботів до адаптивних систем контролю якості.

Актуальність теми зумовлена стрімким зростанням обсягів мультимедійних даних, потребою у високоточних системах розпізнавання для автоматизації виробничих процесів, безпеки, медицини, автономного транспорту та робототехніки. Традиційні одномодальні підходи часто не забезпечують достатньої точності в умовах шуму, неповноти або неоднозначності даних. Мультимодальні системи, навпаки, дозволяють компенсувати обмеження однієї модальності за рахунок іншої, що робить їх більш стійкими та ефективними.

Наукова значущість дослідження полягає у подальшому розвитку методів інтеграції даних різної природи за допомогою нейронних мереж і

технологій глибинного навчання. Практична цінність полягає у можливості покращення завдяки цій роботі у системах відеоспостереження, контролю виробництва, медичній діагностиці, навігації роботів тощо.

Об’єкт дослідження: Процес мультимодального розпізнавання об’єктів за допомогою систем штучного інтелекту.

Предмет дослідження: Моделі, методи та алгоритми обробки, інтеграції та аналізу мультимодальних даних (зображень, аудіо, тексту) для підвищення ефективності систем розпізнавання.

Мета дослідження: Дослідити мультимодальне розпізнавання об’єктів із використанням технологій штучного інтелекту, що забезпечують підвищення точності та надійності класифікації об’єктів.

РОЗДІЛ 1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Загальні основи та сучасний стан розвитку мультимодального розпізнавання об’єктів.

Системи розпізнавання об’єктів становлять одну з фундаментальних задач галузі комп’ютерного зору (computer vision) та штучного інтелекту (ШІ). Під «розпізнаванням об’єктів» зазвичай розуміють задачу, в якій система повинна визначити, чи присутній у зображенні або відеопотоці об’єкт певного класу, а також встановити його положення — наприклад, у вигляді обмежувальної рамки чи контуру. У ширшому сенсі це включає класифікацію (до якого класу належить об’єкт) і локалізацію (де саме він знаходиться).

На практиці системи розпізнавання об’єктів застосовуються в різноманітних сферах — від відеоспостереження, автономного транспорту та медицини до промислової автоматизації й робототехніки (рис 1.1). Це свідчить про те, що вдосконалення таких систем за точністю, швидкодією та надійністю має велике прикладне значення.

Подальший розвиток технологій у цій сфері нерозривно пов’язаний із прогресом у штучному інтелекті. **Штучний інтелект** (Artificial Intelligence) охоплює сукупність методів, що дозволяють створювати системи, здатні до раціонального мислення, навчання, аналізу даних та прийняття рішень. У межах цієї галузі формується машинне навчання (Machine Learning) — підхід, який надає системам можливість самостійно поліпшувати свою роботу на основі аналізу попереднього досвіду, без потреби в жорстко заданих алгоритмах.

Подальшим кроком еволюції є **глибоке навчання** (Deep Learning) — методологія, що базується на використанні штучних нейронних мереж із багатьма рівнями представлення даних. Саме глибоке навчання стало основою революційних досягнень у сфері розпізнавання зображень, мови,

тексту та інших типів інформації. Воно дозволило створювати моделі, які не лише ідентифікують об’єкти, а й “розуміють” контекст сцени, зв’язки між елементами та семантичне значення даних.

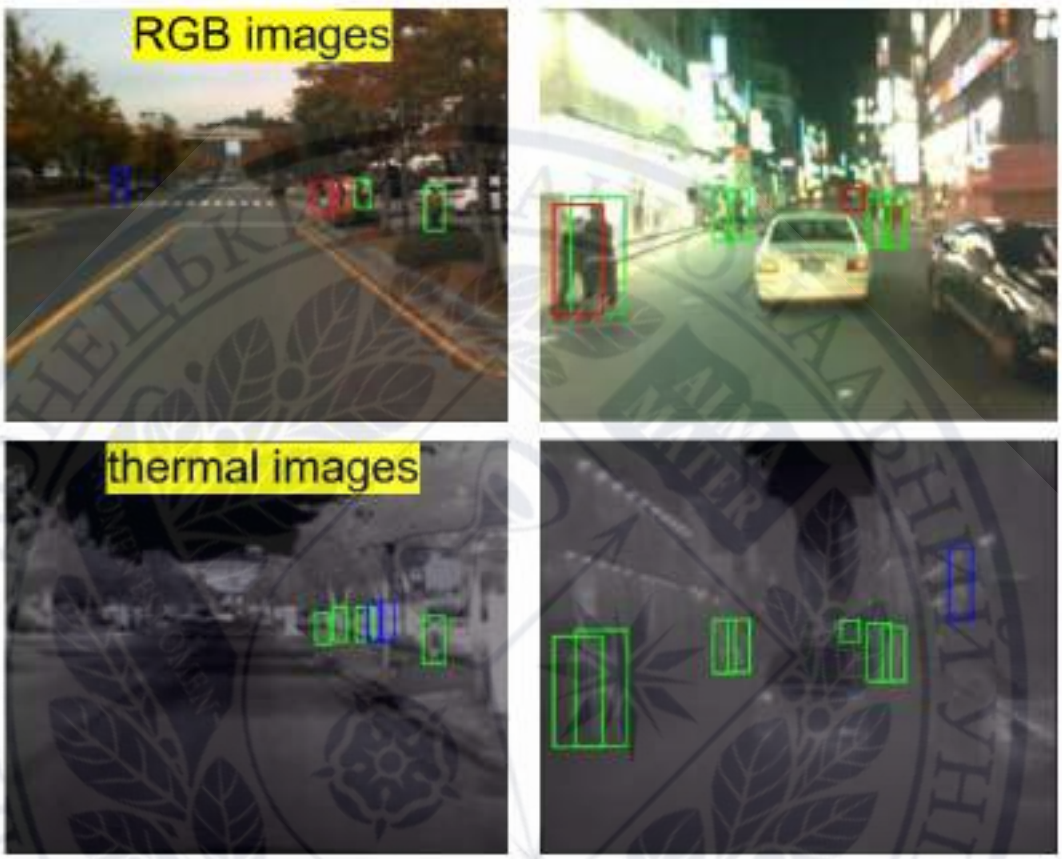

Рисунок 1.1 — Приклад мультимодального розпізнавання об'єктів

Важливою галуззю, у якій поєдналися ці досягнення, є **комп’ютерний зір** (Computer Vision) — напрям штучного інтелекту, що досліджує методи отримання, обробки та інтерпретації візуальної інформації. Метою комп’ютерного зору є навчити машини “бачити” світ подібно до людини, розпізнаючи об’єкти, сцени, дії або закономірності у візуальних даних. Сучасні системи комп’ютерного зору поєднують у собі алгоритми глибокого навчання та великі обсяги різноманітних даних, що робить їх здатними до самонавчання й адаптації до складних умов.

Саме завдяки синтезу штучного інтелекту, машинного та глибокого навчання, а також методів комп'ютерного зору стало можливим створення мультимодальних систем розпізнавання об'єктів. Такі системи здатні об'єднувати інформацію з різних джерел — зображення, тексту, звуку, просторових або сенсорних сигналів — щоб формувати більш повне, узгоджене й осмислене уявлення про навколишнє середовище. Це робить мультимодальний підхід одним із ключових напрямів розвитку сучасного штучного інтелекту.

1.1.1 Мультимодальність і модальність даних

Мультимодальність — це здатність систем штучного інтелекту сприймати, аналізувати та поєднувати інформацію, отриману з різних джерел або модальностей даних. Кожна модальність відображає певний спосіб представлення світу — наприклад, зображення, звук, текст, відео або просторові сигнали. У природному середовищі ці форми інформації взаємопов'язані: люди одночасно бачать, чують і відчувають контекст, завдяки чому формують більш точне розуміння ситуації.

Подібним чином, у штучному інтелекті мультимодальні системи прагнуть відтворити цей принцип — об'єднуючи різні типи даних, вони забезпечують глибше та контекстно багатше сприйняття. Наприклад, система розпізнавання об'єктів може не лише аналізувати зображення, а й враховувати текстові описи або звукові сигнали, що супроводжують сцену. Це дозволяє зменшити неоднозначність і покращити точність прийняття рішень.

Модальність даних, у свою чергу, — це окремий вид або джерело інформації, який характеризується власною структурою та способом сприйняття. Текст передає зміст через мову, зображення — через візуальні ознаки, аудіо — через звукові характеристики, а відео — через поєднання

руху та контексту часу. Вивчення того, як ці модальності взаємодіють між собою, є ключем до створення систем, здатних мислити більш «людиноподібно» — тобто розуміти не лише дані, а й смислові зв’язки між ними.

1) **Джерела даних і типи модальностей**.

Для ефективного мультимодального розпізнавання об’єктів ключовим є розуміння джерел інформації та їх типів. Дані, що надходять у систему, відображають різні аспекти навколишнього середовища, і кожен тип несе унікальну інформацію, що може бути використана для підвищення точності та надійності системи (наглядний приклад можна побачити на рис 1.2).

- **Візуальні дані**

  Візуальні дані, такі як зображення та відео, є основним джерелом інформації у більшості систем комп’ютерного зору. Вони дозволяють аналізувати форму, текстуру, колір, рух та просторові взаємозв’язки об’єктів. Використання відеопотоку дає можливість відстежувати об’єкти в часі та враховувати динаміку сцени. Візуальні дані є фундаментальною модальністю для більшості мультимодальних систем, оскільки вони забезпечують пряме уявлення про фізичні об’єкти та середовище.

- **Текстові дані**

  Текстові дані включають описи, підписи, запити або будь-яку іншу форму словесного представлення інформації. У мультимодальних системах текст часто використовується для уточнення або пояснення візуальної сцени. Наприклад, підписи

до зображень допомагають зв’язати семантичне значення об’єктів з їхнім виглядом, а текстові запити дозволяють здійснювати пошук або керувати системою на природній мові.

- **Аудіо- та мовні дані**

  Аудіо- та мовні дані включають звуки, мову, шумові сигнали або інші аудіопотоки, що можуть супроводжувати візуальну інформацію. Аудіо дозволяє отримати додатковий контекст: наприклад, розпізнавання сигналів руху, голосових команд або джерел звуку у сцені. Комбінування аудіо з іншими модальностями підвищує здатність системи правильно ідентифікувати події та об’єкти.

- **Просторові дані**

  Просторові дані включають інформацію про глибину, тривимірну форму або розташування об’єктів у просторі. Це можуть бути карти глибин (Depth), дані LiDAR або 3D-сцени, отримані за допомогою спеціалізованих сенсорів. Просторові дані дозволяють системам не лише розпізнавати об’єкти, а й оцінювати їхні розміри, положення та взаємне розташування, що критично для автономної навігації, робототехніки та промислової автоматизації.

- **Біометричні дані**

  Біометричні дані включають тепловізійні зображення, рухові сигнали, відбитки пальців, зразки голосу та інші фізіологічні або поведінкові характеристики людини. Такі дані використовуються для ідентифікації, спостереження або контролю доступу, а також для оцінки стану людини у взаємодії з системою. Вони часто доповнюють основні модальності,

забезпечуючи додатковий рівень інформації для підвищення точності розпізнавання.

Рисунок 1.2 — Тип модальності, джерело та вигляд даних

2) Сенсорний ф’южн та інтеграція модальностей

Отже, ми розглянули основні типи даних, які використовуються в мультимодальних системах розпізнавання — від візуальних і текстових до аудіо, просторових і біометричних. Проте саме по собі збирання таких даних ще не забезпечує розуміння чи точного аналізу. Щоб отримати цілісну картину, необхідно навчитися поєднувати ці різні потоки інформації — узгоджувати їх між собою, виявляти спільні закономірності та доповнювати недоліки однієї модальності перевагами іншої. Саме цей процес — **сенсорний ф’южн** та **інтеграція модальностей** — лежить в основі ефективного мультимодального розпізнавання.

Під **сенсорним ф’южном** розуміють узгодження інформації, отриманої з кількох сенсорів або джерел даних (рис 1.3). Метою є формування більш надійного, повного й узагальненого представлення навколишнього середовища. Наприклад, камера може надати візуальне зображення об’єкта, тоді як LiDAR — його просторову форму, а мікрофон — аудіоконтекст. Об’єднання цих сигналів дозволяє системі приймати рішення з вищою точністю, ніж при використанні одного джерела.

Рисунок 1.3 — Схематичний приклад сенсорного ф’южн

У ширшому сенсі, **інтеграція модальностей** охоплює не лише технічне об’єднання сигналів із сенсорів, а й когнітивний рівень — узгодження різних способів сприйняття: зору, мови, звуку, тексту, простору. Такі системи не просто «зливають» дані, а намагаються знайти між ними спільне семантичне значення — наприклад, співвіднести опис “червона машина” з конкретним візуальним об’єктом на зображенні [1].

Процес ф’южну зазвичай описують через три основні рівні інтеграції (схематичний вигляд можна побачити на (рис 1.4):

1. **Рівень даних (Data-level fusion) або Early fusion** — поєднання сирих або мінімально оброблених даних з різних сенсорів. Наприклад,

комбінування глибинної карти та зображення RGB у єдину просторову модель.

2. **Рівень ознак (Feature-level fusion) або Hybrid fusion** — інтеграція вже оброблених або витягнутих характеристик, таких як контури, текстури, вектори ознак. Цей підхід часто дозволяє зменшити обсяг даних, але зберегти їх змістовність.
3. **Рівень рішень (Decision-level fusion) або Late fusion** — поєднання результатів окремих моделей або класифікаторів у єдине підсумкове рішення. Наприклад, якщо модель з обробки зображень вважає, що об’єкт — автомобіль, а модель на основі звуку двигуна це підтверджує, система робить спільний висновок з більшою впевненістю [2].

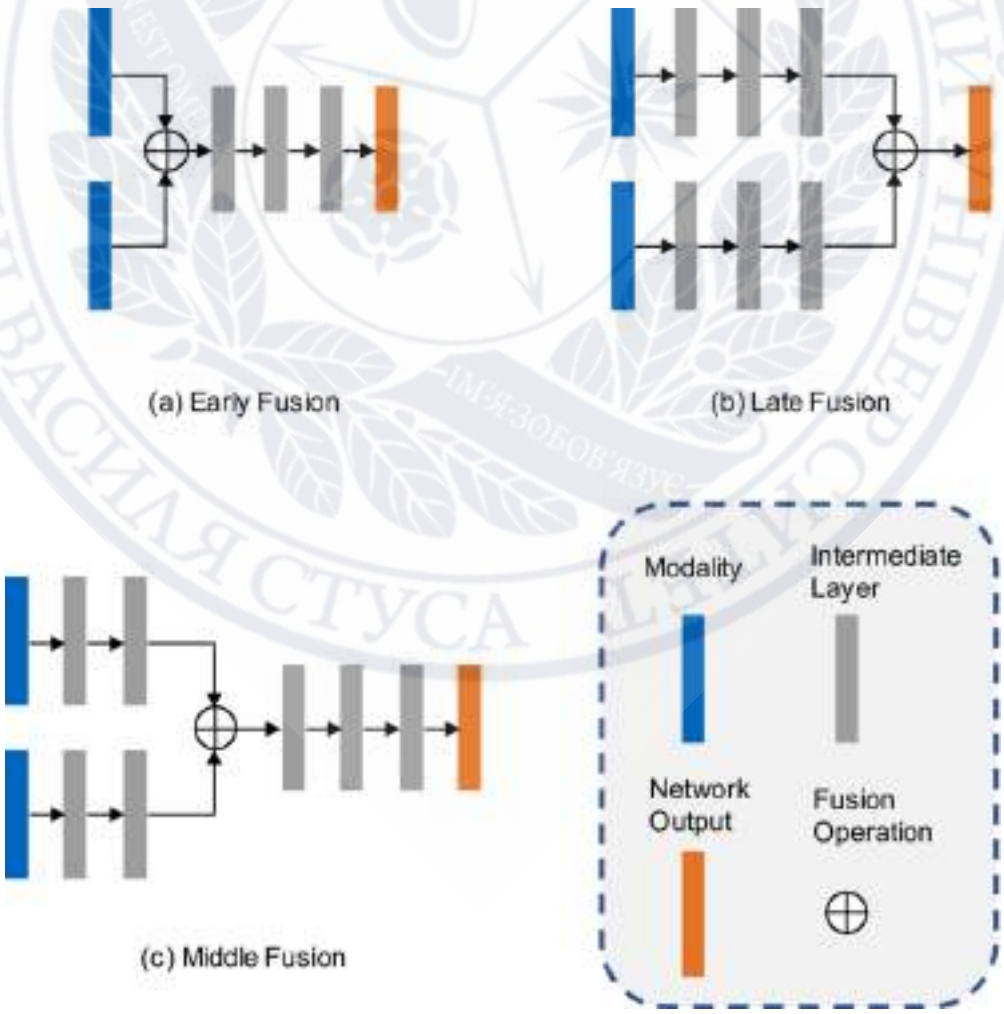

Рисунок 1.4 — Рівні злиття для мультимодального виявлення об'єктів

У сучасних мультимодальних системах часто поєднують кілька рівнів ф’южну — наприклад, первинне об’єднання даних на рівні ознак і подальше уточнення на рівні рішень. Такий підхід забезпечує баланс між гнучкістю, точністю та ефективністю обробки.

Таким чином, сенсорний ф’южн та інтеграція модальностей є ключовими елементами, що дозволяють системам штучного інтелекту сприймати світ більш «людяно» — не лише бачити, а й розуміти контекст, пов’язуючи різні джерела інформації в єдину, осмислену структуру.

1.1.2 Ключові виклики при створенні мультимодальних систем.

Розробка мультимодальних систем штучного інтелекту супроводжується низкою суттєвих проблем, що зумовлені необхідністю інтеграції різнорідних типів даних, забезпеченням стабільної продуктивності та керуванням зростаючою обчислювальною складністю. Такі системи мають одночасно обробляти дані різної природи — текстові, візуальні, аудіальні та сенсорні — що потребує врахування особливостей кожного типу інформації, способів її подання, узгодження та синхронізації.

1) Інтеграція та вирівнювання даних різних модальностей:

Однією з ключових проблем є необхідність гармонійного поєднання даних із різних джерел. Текстові дані є послідовними та символічними, тоді як зображення мають просторову структуру і складаються з пікселів. Для ефективної обробки цих відмінностей використовуються спеціалізовані архітектури — згорткові нейронні мережі для візуальних даних та трансформери для тексту. Головне завдання полягає у створенні механізмів, які забезпечують змістовні взаємозв’язки між цими модальностями, дозволяючи моделі інтегрувати різні типи інформації у єдиному контексті.

Додатковою складністю є часове вирівнювання даних. Наприклад, у відеоаналізі потрібно забезпечити синхронізацію звукових сигналів із відповідними візуальними кадрами. Навіть незначне відхилення у вирівнюванні може призвести до зниження точності системи або хибних інтерпретацій подій.

2) Робота з неповними та зашумленими даними:

У реальних умовах мультимодальні системи часто мають справу з неповними або неякісними даними. У багатьох наборах інформації відсутні певні компоненти — наприклад, бракує аудіо або зображень, що порушує цілісність модальностей. Для подолання цього застосовуються методи міжмодального перенесення знань, коли інформація з однієї модальності допомагає заповнити прогалини в іншій. Так, якщо система генерації підписів до зображень стикається з недостатньо позначеним візуальним прикладом, вона може використати знання, набуті з інших, добре розмічених даних.

Ще однією серйозною проблемою є наявність шуму, наприклад, фонового звуку в аудіо або розмиття руху у відео. Для підвищення стійкості до таких спотворень необхідно створювати складні конвеєри попередньої обробки, що значно ускладнює архітектуру системи та підвищує обчислювальні витрати.

Крім того, упередження в одній модальності можуть передаватися іншим. Наприклад, якщо текстові дані містять стереотипні або нерепрезентативні патерни, це може спричинити викривлення в комбінованих результатах, знижуючи достовірність висновків системи.

3) Обчислювальні ресурси та масштабованість:

Ще один вагомий виклик — забезпечення ефективності та масштабованості мультимодальних моделей. Такі системи зазвичай

складаються з кількох паралельних нейронних мереж, кожна з яких спеціалізується на певному типі даних. Це призводить до високих вимог щодо пам’яті та обчислювальної потужності. Навчання подібних архітектур потребує спеціалізованого апаратного забезпечення — графічних процесорів (GPU) або тензорних процесорів (TPU) — що обмежує можливості їх використання невеликими дослідницькими командами чи стартапами.

Розгортання мультимодальних систем на периферійних пристроях (наприклад, смартфонах чи вбудованих сенсорах) вимагає застосування методів оптимізації: обрізання моделі, квантування або стискання параметрів. Проте такі підходи часто знижують точність моделі. Для прикладу, мобільний додаток перекладу в реальному часі, який комбінує голосову та текстову інформацію, повинен знайти баланс між швидкодією та точністю, жертвуючи одним аспектом заради іншого.

4) Продуктивність у реальних умовах:

Підтримання стабільної продуктивності мультимодальних систем у різних апаратних та середовищних умовах залишається відкритою проблемою. Необхідність компромісу між ефективністю, якістю обробки та ресурсними обмеженнями ускладнює проєктування універсальних рішень. Розв’язання цих питань потребує подальших досліджень у напрямах оптимізації архітектур, динамічного балансування навантаження та адаптивного керування ресурсами [3].

1.1.3 Розвиток методів розпізнавання об’єктів за допомогою ШІ

Розвиток методів розпізнавання об’єктів можна умовно розділити на два великі етапи: **традиційні алгоритмічні підходи** і **епоху глибокого навчання** (deep learning).

На ранніх етапах (1990-ті – початок 2000-х) застосовувалися методи, засновані на виділенні ознак (features) — наприклад, детектори країв, контурів, дескриптори типу SIFT (Scale-Invariant Feature Transform), SURF (Speeded Up Robust Features), HOG (Histogram of Oriented Gradients), Haar-подібні ознаки тощо. Системи будувалися за класичною схемою: виділення ознак → класифікація → (за потреби) локалізація. Проте ці методи мали суттєві обмеження: чутливість до зміни освітлення, масштабу, повороту, фону, заокруглень, часткового закриття об’єктів (occlusion) .

З початком застосування нейронних мереж і особливо глибоких згорткових мереж (CNN – convolutional neural networks) відбувся якісний науковий прорив. У праці Girshick було запропоновано підхід R-CNN, який започаткував нову еру — автоматичне навчання ознак з великою кількістю даних. В подальшому з’явилися Faster R-CNN, SSD (Single Shot MultiBox Detector), YOLO (You Only Look Once) — котрі поєднували швидкість і точність [4].

Останнім часом, починаючи приблизно з 2020-х років, у моделях розпізнавання об’єктів активно застосовуються архітектури на базі **Transformer**, що дозволяють краще враховувати глобальні залежності у зображенні та контексти. Модель на основі трансформера DEtection TRansformer (DETR), започаткувала нову еру в обробці зображень, перетворивши задачу детекції на задачу передбачення множини об'єктів без необхідності попереднього генерування пропозицій чи постобробки, що значно спростило архітектуру системи [5].

Відповідно на час 2020 ого року, системи вже мають високу продуктивність, можуть працювати в реальному часі, але все ще стикаються з проблемами (наприклад невдосконалена мультимодальність, недостатньо даних або обмеження ресурсів).

З 2023 року також спостерігається активний розвиток трансформерів для детекції дрібних об'єктів. Дослідження показали, що трансформери переважають традиційні CNN-методи в задачах детекції дрібних об'єктів,

таких як виявлення об'єктів у відео та зображеннях, завдяки своїй здатності моделювати глобальні залежності та контекстуальну інформацію [6].

Останні досягнення в галузі включають розвиток моделей, що поєднують трансформери з іншими підходами, такими як Vision-Language Models (VLMs), для покращення детекції в 3D-просторі. Наприклад, Quantum Inverse Contextual Vision Transformers (Q-ICVT) використовують квантові концепції для інтеграції даних з різних сенсорів, що дозволяє досягти високої точності детекції в автономних транспортних засобах [7].

Сучасний етап характеризується переходом від ручного конструювання ознак до автоматичного навчання, від локального аналізу до глобальних моделей, що враховують як контекст, так і модальності.

1.1.4 Етичні проблеми в мультимодальних системах штучного інтелекту.

Мультимодальні системи штучного інтелекту, що інтегрують різні типи даних — текст, зображення, відео чи аудіо, — відкривають нові можливості, але водночас породжують складні етичні дилеми. Найбільш суттєвими серед них є три проблеми: упередженість та нерівність, ризики порушення конфіденційності і нестача прозорості та підзвітності. Поєднання різнорідних джерел даних значно ускладнює контроль за тим, як саме формується рішення системи, і вимагає від розробників відповідального підходу до її проєктування та розгортання.

1) Упередженість і справедливість:

Мультимодальні моделі мають тенденцію відтворювати та підсилювати упередження, що вже закладені в окремих модальностях даних. Наприклад, система відбору персоналу, яка аналізує одночасно текстові резюме та відеозаписи співбесід, може виявитися схильною до

дискримінації. Текстовий аналізатор може віддавати перевагу певним формулюванням, які частіше використовують представники конкретних соціальних груп, а алгоритм розпізнавання облич — помилятися при ідентифікації людей з темним відтінком шкіри. Таке поєднання помилок призводить до несправедливих рішень і поглиблення соціальної нерівності.

Крім того, недостатнє представлення меншин у навчальних наборах зображень чи відео створює систематичні викривлення. Для мінімізації таких ефектів необхідні ретельні аудити даних усіх модальностей, тестування результатів на справедливість у межах різних демографічних груп і застосування методів корекції упередженості. Інакше система може несвідомо виключати кваліфікованих кандидатів через акцент, зовнішність або інші несуттєві чинники.

2) Конфіденційність і захист даних:

Збирання та обробка кількох типів інформації значно підвищує ризики витоку приватних даних. Наприклад, медичний застосунок, який одночасно оперує історією хвороби, зображеннями МРТ та голосовими нотатками лікаря, має гарантувати безпеку кожного каналу окремо. Компрометація хоча б одного типу даних може автоматично поставити під загрозу інші.

До того ж мультимодальні алгоритми здатні виводити неочевидну інформацію — наприклад, поєднання часових міток із геолокацією у фото може розкрити повсякденні маршрути користувача. Щоб запобігти цьому, розробники повинні застосовувати принципи мінімізації даних, обмеження доступу та анонімізації. Так, фітнес-додаток, який використовує голосові та рухові сигнали, не повинен зберігати необроблені записи, якщо в цьому немає прямої потреби. Інакше існує ризик порушення норм на кшталт GDPR та втрати довіри користувачів.

3) Прозорість і підзвітність:

Однією з найбільших труднощів є пояснення процесу ухвалення рішень у мультимодальних системах. Коли інструмент виявлення шахрайства формує попередження на основі текстових повідомлень і геолокаційних даних, стає незрозуміло, який саме канал вплинув на підсумкове рішення. Без чітких пояснень користувачі не можуть оскаржити помилкові результати, а розробники — визначити джерело проблеми.

Ця непрозорість створює труднощі й для визначення відповідальності. Якщо автономний автомобіль потрапив в аварію через розбіжності між даними камери та лідара, постає питання: хто винен — виробник датчика чи команда, що створювала алгоритм обробки? Для уникнення таких ситуацій рекомендується впроваджувати модульну архітектуру системи, вести журнали аудиту процесу прийняття рішень і створювати детальну документацію, яка чітко описує роль кожної модальності.

Прозоре пояснення впливу різних типів даних, наприклад при оцінюванні кредитного рейтингу, підвищує підзвітність системи, сприяє дотриманню правових вимог і зміцнює довіру користувачів до технологій штучного інтелекту [8].

1.2 Огляд літератури

Так як розвиток мультимодального розпізнавання об’єктів напряму пов’язаний та залежить від наукових публікацій пов’язаних із одномодальним розпізнавання об’єктів, буде розглянуто також і сферу розпізнавання об’єктів загалом за останні роки, а також їх взаємозв’язок із темою мультимодального розпізнавання. Нижче наведено основні напрями, з аналізом, порівняннями та викликами для науковців.

1.2.1 Ключові напрями сучасних досліджень у загальній сфері розпізнавання об’єктів і їх взаємозв’язок із мультимодальним розпізнаванням:

1) Оптимізація архітектур мереж для більшої продуктивності

Один із ключових напрямів — це постійне вдосконалення архітектур нейронних мереж, використаних для розпізнавання об’єктів. Сучасні архітектури нейронних мереж (YOLOv8, EfficientDet, DETR, ConvNeXt тощо) спрямовані на підвищення продуктивності, швидкості обробки та зниження споживання ресурсів. Вони оптимізуються під edge-пристрої, мобільні системи та сценарії з обмеженими обчислювальними можливостями.

У аспекти вдосконалення архітектур нейромереж входять: збільшення глибини і ширини мереж, впровадження attention-механізмів, комбінування класичних CNN із Transformer-компонентами. Наприклад, в огляді «A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection» зазначено, що сучасні моделі переходять до поєднання різних підходів поєднання CNN із Transformer, щоб досягнути балансу точності та швидкодії [9].

Також, архітектурні напрямки типу «one-stage» vs «two-stage» ще продовжують бути предметом досліджень: двоетапні підходи (наприклад, R-CNN-подібні) забезпечують вищу точність, але мають вищу складність; одноетапні рішення (наприклад, YOLO, SSD) — більш швидкі, але іноді менш точні [10]. Одноетапні моделі краще підходять для реального часу, мобільних/edge сценаріїв, тоді як двоетапні — для задач із високими вимогами до точності. Тож ключовий виклик для науковців це взнати, як зберегти високу точність при зменшенні обчислювальних ресурсів.

У мультимодальному контексті оптимізація архітектури відіграє ключову роль, адже інтеграція кількох каналів (RGB, глибина, LiDAR, аудіо, текст тощо) потребує значно більшої кількості параметрів і

обчислень. Збалансованість між точністю та швидкодією тут є вирішальною — і саме оптимізовані моделі дозволяють мультимодальним системам працювати в реальному часі.

Наприклад: Оптимізовані архітектури полегшують fusion (злиття) модальностей і знижують затримку в реальному часі. Оптимізовані моделі дозволяють створювати легкі мультимодальні моделі для edge-платформ. Також краще підтримують асинхронне об’єднання модальностей, коли не всі канали доступні одночасно. EfficientDet та MobileNet використовуються як базові енкодери для мультимодальних моделей злиття RGB + Depth [11]. Архітектура Transformer-типу DETR адаптована для багатомодального навчання, завдяки універсальному механізму attention [12].

2) Виявлення та розпізнавання об’єктів у не контрольованих умовах

Інший значущий напрям — адаптація моделей до реальних, складних умов: малий розмір об’єкта, мала вибірка, сильне перекривання об’єкта (occlusion), змінювальне освітлення, складний або нестабільний фон. У статті «Survey and Performance Analysis of Deep Learning Based Object Detection in Challenging Environments» описано і проаналізовано підходи до покращення. В статті описана як продуктивність моделей падає в таких складних, і як ще кажуть, не контрольованих умовах, [13]. Ці проблеми стимулюють дослідників розробляти нові методи попередньої обробки, підсилення даних за допомогою нарощування даних (data augmentation) - штучного збільшення вибірки для навчання трансформуючи оригінальні дані [14], спеціалізовані збірки даних, а також мережі з навчанням за слабшою розміткою (weakly-supervised) чи без розмітки (self-supervised) [15] .

Якщо порівнювати моделі які були розроблені лише на, або для, контрольованих умов, то вони часто демонструють значно гірші результати в реальному застосуванні ніж ті які пройшли калібрування і перенавчання у

реальних умовах. Таким чином, цей напрям — критично важливий для прикладних задач (автономні системи, дрони, відеоспостереження). Тобто загальна мета досліджень у цьому напрямку, це створення **універсальної моделі** — архітектури, що стабільно працює у «поганих» умовах, це в свою чергу так само стосується і **мультимодальних моделей**.

3) Мультимодальність на edge-платформах

З ростом застосувань у мобільних пристроях, IoT, автономних системах з'явився потужний напрям — розробка «легких» моделей (lightweight models), які займають менше пам'яті, мають нижчу затримку, менш вимогливі до апаратури. Наприклад, в оглядах вказано, що навіть великі моделі тягнуть значні ресурси, тому створення нових варіантів реалізації та архітектур для edge-пристроїв — ключовий напрямок розвитку [9]. Для кращого розуміння **Edge-пристрої** — це обчислювальні пристрої, які знаходяться ближче до джерела даних або користувача (локальні) і виконують обробку, зберігання або аналіз даних, без необхідності постійного звернення до хмарного центру.

Зріст оптимізації дає можливість впровадження мультимодальних систем розпізнавання об'єктів у реальних прикладних сценаріях (розумні міста, автономний транспорт, відеоспостереження, роботи) та на обладнанні з обмеженими ресурсами. Наприклад, дослідження «Multimodal Object Detection: An Architecture Using Feature-Level Fusion and Deep Learning» (Neural Computing and Applications, 2025) описує оптимізовану архітектуру для edge-сценарію, де багато сенсорів на вуличній інфраструктурі збирають дані. [16].

У порівнянні, класичні системи розпізнавання об'єктів часто були серверними, із великою потужністю, а edge-версії мають компроміси — менш точні, але більш швидкі і доступні. Наукові роботи за цією темою намагаються вирішити такі проблеми: як забезпечити енергетичну

ефективність — через квантізацію, прунінг, оптимізовані інфраструктури, узгодженість модальностей і зменшення затримки в реальному часі.

Усі ці напрями наукових досліджень взаємопов'язані: наприклад, мультимодальність і переносимість часто підсилюють один одного; оптимізація архітектур значно важливіша для edge-моделей, але розвиток оптимізації так само позитивно впливає і на великі, потужні, мультимодальні моделі.

Загалом у розвитку систем розпізнавання об'єктів простежується чітка тенденція до зростання складності, інтелектуальності та мультимодальності. Сучасні дослідження не лише вдосконалюють алгоритми класифікації чи детекції, а й оптимізують продуктивність моделей, підвищують їх адаптивність до різних умов використання, покращують інтеграцію різних модальностей та забезпечують більш ефективне використання обчислювальних ресурсів. Це створює підґрунтя для появи нових інтелектуальних систем, здатних до гнучкої взаємодії з навколишнім середовищем та людиною, що є ключовим кроком у розвитку технологій штучного інтелекту.

1.2.2 Напрями розвитку мультимодальних систем розпізнавання об'єктів

1) Об'єднання зображень із текстом та мовою (Vision-Language)

Одним із найперспективніших напрямів сучасних досліджень є інтеграція візуальної інформації з текстовими та мовними модальностями (рис 1.5). Такі моделі — як CLIP (Contrastive Language–Image Pre-training), Flamingo, LLaVA чи Kosmos — здатні обробляти зображення й текст одночасно, створюючи спільний латентний простір, у якому відображення обох модальностей узгоджуються за змістом [17].

Це дозволяє не лише визначати, що зображено на фото, а й розуміти, як опис користувача співвідноситься із зображенням. Наприклад, система може виконати пошук за запитом “знайди на фото предмет, який нагадує антену” — ідентифікуючи об’єкт навіть без попереднього навчання саме на цій категорії.

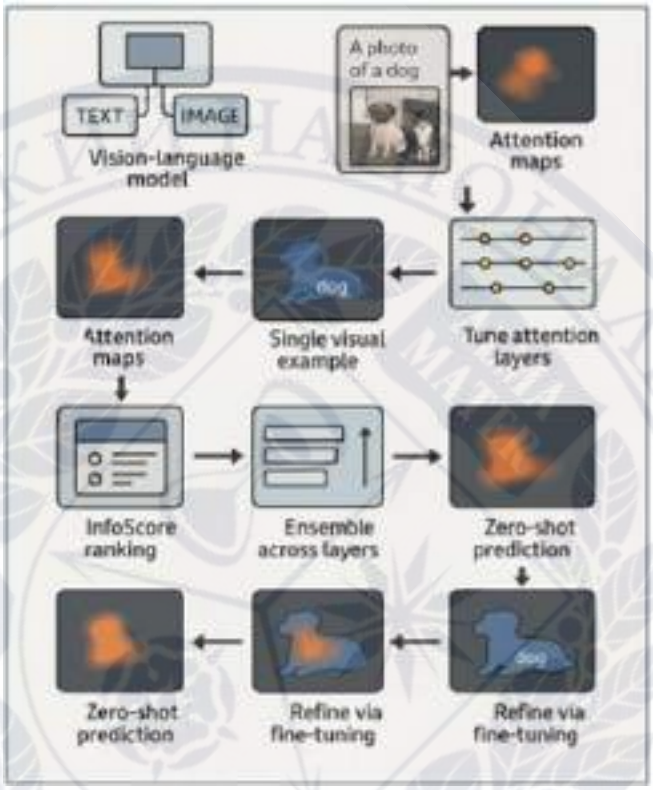

Рисунок 1.5 — Схематичний приклад роботи **Vision-Language моделі**

У сучасних дослідженнях підкреслюється, що мультимодальні великі моделі розширюють межі детекції, адже вони поєднують семантичне розуміння тексту з візуальним контекстом [4]. Подібні роботи також показують, що системи типу Grounding DINO чи OWL-ViT демонструють успішну інтеграцію між візуальними та мовними описами, дозволяючи виконувати детекцію об’єктів із відкритим словником (open-vocabulary detection) [18]. Така мультимодальність є логічним розвитком традиційних моделей: якщо звичайна система просто визначає що це за об’єкт, то мультимодальна система може відповісти який саме об’єкт відповідає

певному опису. Це підвищує рівень гнучкості та пояснюваності моделей, особливо у випадках, коли візуальна інформація неоднозначна. Проте, залишається складність узгодження різнорідних представлень: надмірна вага текстової підказки може спричинити хибну інтерпретацію візуальної сцени, тоді як слабке текстове моделювання знижує якість мультимодальної інтеграції.

2) Покращені стратегії ф'южну (fusion) та боротьба з "fusion degradation"

Напрям, який досліджує, як правильно поєднувати різні типи даних (RGB-зображення, глибину, текст, термальні сигнали тощо) в межах однієї мультимодальної системи, щоб отримати повнішу та точнішу інтерпретацію сцени. Основна ідея полягає в тому, що кожна модальність містить унікальну інформацію і якщо їх об'єднати узгоджено, модель отримує більш глибоке розуміння об'єкта чи сцени.

Проблема «fusion degradation» виникає, коли неправильне або надмірне поєднання призводить до втрати корисної інформації, шуму чи суперечливих сигналів між модальностями. Це може знижувати точність і стабільність моделі [19].

Тому сучасні дослідження фокусуються на **легких, гнучких і адаптивних схемах ф'южну**, які динамічно визначають, яку модальність коли й у якому обсязі враховувати. Використовуються різні методи: вагове злиття на основі уваги (attention-based fusion), крос-модальні трансформери, а також механізми вибіркового ф'южну, які автоматично пригнічують шум і підсилюють найінформативніші ознаки.

Завдяки цьому підходи нового покоління забезпечують **стійкі результати навіть у складних умовах** — наприклад, при зміні освітлення, відсутності кольору чи часткових спотвореннях, — і відкривають шлях до створення ефективних мультимодальних систем розпізнавання об'єктів. Головна задача науковців зрозуміти як організувати навчання, коли дані

окремих модальностей різної якості? Або як уникнути «домінування» однієї модальності?

3) Мультимодальна 3D-детекція

Мультимодальна 3D-детекція — це сучасний напрям, у якому поєднуються різні джерела просторової та семантичної інформації (наприклад, RGB-зображення, дані LiDAR, карти глибини чи 3D-атрибути) для точнішого розпізнавання об’єктів у тривимірному просторі.

Завдяки комбінуванню цих модальностей модель отримує не лише візуальні ознаки, а й розуміння форми, розташування та просторової структури об’єктів. Це особливо важливо для робототехніки та автономних систем, які повинні орієнтуватися в реальному світі, приймати рішення та взаємодіяти з оточенням. [20].

Останнім часом швидко зростає інтерес до **open-vocabulary 3D-детекції**, коли система може розпізнавати об’єкти **поза фіксованим набором класів**, використовуючи текстові підказки чи описові запити. Такі моделі здатні інтегрувати **мовні представлення у 3D-сцени**, описуючи не лише тип об’єкта, а й його властивості, контекст або взаємодії з іншими елементами. У поєднанні з візуально-мовними моделями це відкриває перспективи створення **універсальних 3D-систем сприйняття**, які можуть гнучко розуміти та інтерпретувати складні сцени без потреби в жорстко визначених класах або ручній розмітці. [21].

4) MLLM / LVLM для локалізації й управління (grounding, robotics, embodied AI)

Мультимодальні великі мовні моделі (MLLM — Multimodal Large Language Models, або LVLM — Large Models) дедалі частіше розглядаються як “мозок” для систем, які поєднують сприйняття, розуміння та дію — зокрема в задачах локалізації, управління роботами та створення “втіленого ШІ” (Embodied AI). Їх роль полягає у здатності інтегрувати знання з різних

модальностей — зображень, відео, 3D-сцен, сенсорних даних, текстових команд — і перетворювати їх на осмислені дії або рішення.

У контексті **локалізації** MLLM можуть не лише описувати об’єкти, а й “прив’язувати” текстові поняття до конкретних частин візуальної або просторової сцени — процес, який називається *grounding*. Наприклад, коли користувач дає команду “візьми червону чашку з правого столу”, модель не просто розпізнає об’єкти, а пов’язує мовний опис із координатами в просторі. Це поєднання тексту з геометричною інформацією є ключовим для автономних систем, які повинні орієнтуватися у фізичному середовищі.

У сфері **робототехніки** такі моделі дають змогу створювати більш гнучких і розуміючих агентів, які можуть виконувати інструкції, адаптуватися до нових ситуацій і навчатися від природної мови без жорсткого програмування. Завдяки мультимодальним архітектурам MLLM можуть обробляти дані з камер (RGB, depth, LiDAR), мікрофонів і тактильних сенсорів, об’єднуючи їх у спільний семантичний простір. Це дозволяє реалізувати високорівневе керування, де робот не просто реагує на дані, а “розуміє контекст”.

У **Embodied AI** (штучний інтелект, інтегрований у фізичні системи, які можуть сприймати, діяти та навчатися в реальному світі) MLLM стають центральним компонентом, що поєднує сприйняття, планування та дію. Такі системи можуть не лише розпізнавати об’єкти, а й планувати послідовність дій для досягнення цілей. Прикладом є поєднання моделей типу GPT-4V, Flamingo чи OpenVLA (Open Vision-Language-Action) із середовищами симуляції, де вони керують агентами у віртуальних або реальних просторах. Цей напрям швидко розвивається завдяки інтеграції LLM із візуально-просторовими модулями, що робить можливим “інструктивне навчання” роботів через природну мову. Прикладом останніх інновацій та технологічного прогрессу сфері Embodied AI є робот «Амека» (humanoid robot Ameca) (рис 1.6). Ameca в першу чергу розроблена як платформа для подальшого розвитку технологій взаємодії людини та робота. Особлива

сила системи полягає в інтеграції передових моделей штучного інтелекту. GPT-3 від Openai використовувався в попередніх версіях, тоді як новіші моделі працюють з GPT-4O. Ця інтеграція штучного інтелекту дозволяє Ameca плавно говорити на різні теми та зв'язно й продумано відповідати на запитання. У Музеї майбутнього Нюрнберга було розроблено спеціальну систему, яка може імітувати різні «настрої» та риси характеру — від корисної допомоги до іронічних чи гумористичних взаємодій [22].

Рисунок 1.6 — Робот «Амека» (humanoid robot Ameca)

Загалом, MLLM/LVLM поступово перетворюються з моделей для опису зображень на системи загального призначення, здатні “мислити” мультимодально. Вони не лише відповідають на питання про те, *що* зображено, але й розуміють *де* це розташовано, *як* з цим взаємодіяти і *чому* дія має сенс у даному контексті. Таким чином, вони формують основу для нової хвилі універсальних систем сприйняття та управління, які можуть адаптуватися до нових середовищ і завдань без потреби в ручному перепрограмуванні. [23].

Аналіз сучасної наукової літератури показує, що розвиток систем розпізнавання об’єктів відбувається у напрямі інтеграції різних підходів, технологій і модальностей, що поступово формує основу для створення універсальних мультимодальних моделей. Ключові дослідження зосереджені на підвищенні ефективності архітектур нейронних мереж, розширенні можливостей розпізнавання у неконтрольованих середовищах і перенесенні інтелектуальних алгоритмів на edge-платформи.

1.3 Огляд готових інструментів

Через широке застосування цієї технології огляд існуючих рішень охоплює системи різного призначення — від промислового контролю якості до інтерактивних мовних моделей, що поєднують зорову, слухову та текстову модальності. Незважаючи на різницю у функціональності, усі вони базуються на єдиній ідеї — інтеграції даних із різних джерел у спільний векторний простір ознак.

1.3.1 Інструменти штучного інтелекту для розробників мультимодальних систем розпізнавання об’єктів.

Сучасні розробники мають доступ до низки мультимодальних інструментів штучного інтелекту, що дозволяють обробляти та інтегрувати різні типи даних — текст, зображення, аудіо та відео. Такі інструменти здебільшого надаються у вигляді API, бібліотек або фреймворків, що спрощують їх використання у практичних проектах. Від 2022 року під час зростання попиту та популярності LVLM, було створено та оновлено величезна кількість мультимодальних моделей, усі вони стрімко технічно

покращувались, і більшість старих моделей залишилась позаду за незначний проміжок часу через велику конкуренцію та прорив у маштабованості сфери.

Ці досягнення зумовлені появою послідовної серії моделей, кожна з яких вносить інновації у механізми візуальної релевантності, мовно-візуального узгодження та мультимодального мислення (На рисунку 1.7 подано узагальнену хронологію розвитку Large Vision-Language Models (LVLM), орієнтованих на завдання виявлення основних об’єктів, створених у період з 2022 року до сьогодні). Вона демонструє швидку еволюцію цієї галузі та поступове зростання як складності архітектур моделей, так і широти їхніх можливостей. Такий прогрес свідчить про чітку тенденцію до формування уніфікованих мультимодальних представлень, здатних підтримувати широкий спектр завдань — від детекції до сегментації, автоматичної генерації підписів та візуального запитання-відповіді [23].

Монополісти створюють дешеві та якісні токіни, неймовірно швидко підвищивши технічний та науковий прогрес за рахунок попиту та величезних фінансових вкладень. Серед найвідоміших прикладів можна виділити GPT-5 з Vision (GPT-4V) від OpenAI, Gemini від Google та ImageBind від Meta. Кожен із них підтримує унікальні типи вхідних даних та має різні функціональні можливості, що робить їх корисними для створення мультимодальних систем різного призначення.

До створення GPT-5 була створена GPT-4V від OpenAI і яка була розширенням моделі GPT-4, яке додавала можливості аналізу зображень. Зараз GPT-5 має вбудавану Vision-частину [https://blog.roboflow.com/gpt-5-vision-multimodal-evaluation/]. Завдяки API розробники можуть створювати програми, що приймають як текстові, так і візуальні вхідні дані — наприклад, системи, які генерують описи до фотографій або відповідають на запитання про їхній зміст. Прикладом практичного застосування може бути інструмент, що аналізує завантажену користувачем діаграму та надає технічні пояснення.

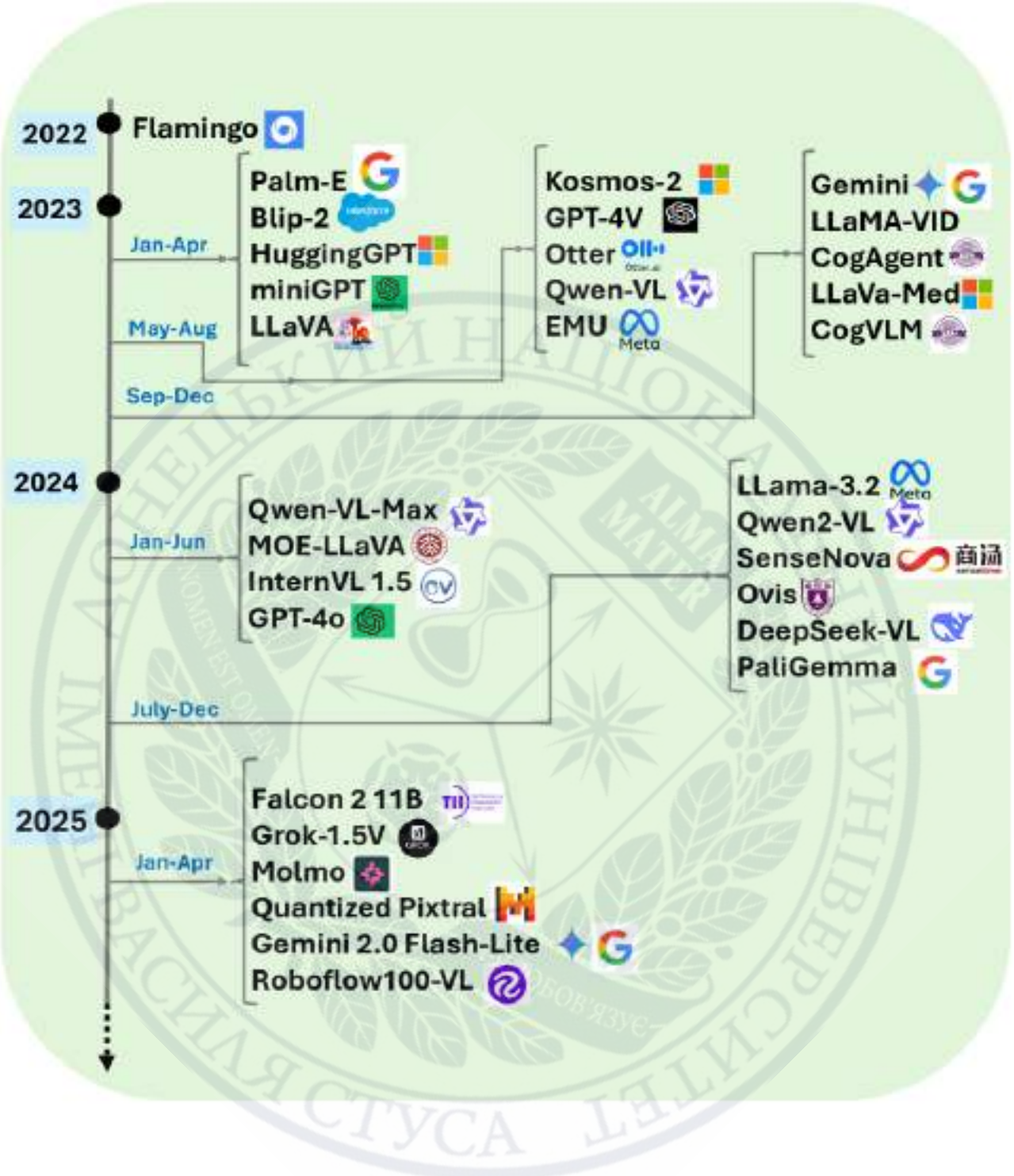

Рисунок 1.7 — Хронологія розвитку Large Vision-Language Models (LVLM)

Gemini від Google є ще більш універсальним рішенням, оскільки здатен безпосередньо обробляти текст, зображення, аудіо та відео. Він надає єдиний API для завдань, що вимагають комплексної обробки даних, зокрема

узагальнення відеоконтенту через поєднання розпізнавання мовлення та аналізу зображень. Це робить його особливо корисним для створення систем автоматичних субтитрів, медіааналізу або модерації відеоконтенту.

ImageBind від Meta є відкритим фреймворком, що поєднує одразу шість типів даних: текст, зображення, аудіо, глибину, теплові карти та IMU-дані (дані інерціальних датчиків). На відміну від більшості інструментів, зосереджених лише на текстово-візуальній взаємодії, ImageBind дозволяє експериментувати з менш поширеними модальностями. Наприклад, можливо створити модель, яка ідентифікує відповідне зображення за звуками навколишнього середовища.

Крім того, бібліотека Transformers від Hugging Face пропонує набір попередньо навчених мультимодальних моделей, серед яких CLIP (поєднує текст і зображення) та Flava (інтегрує текст, зображення та метадані). Ці інструменти легко інтегруються через Python API, що дає змогу швидко створювати прототипи мультимодальних систем без необхідності розробки складних конвеєрів з нуля.

Таким чином, доступність цих рішень значно спрощує розробку сучасних мультимодальних систем штучного інтелекту, дозволяючи фокусуватися на прикладних аспектах інтеграції даних, а не на побудові архітектури з нуля [25].

1.3.2 Мультимодальні інструменти для виробництва.

Одним із прикладів практичного застосування мультимодального розпізнавання є системи контролю дефектів на виробництві. У галузі контролю якості мультимодальний штучний інтелект виділяється як революційна сила. Виробники все частіше інтегрують мультимодальні системи для контролю якості на основі штучного інтелекту, щоб забезпечити відповідність стандартам продукції, зменшити кількість

дефектів та пришвидшити прийняття рішень у режимі реального часу. Ця потужна форма штучного інтелекту змінює процес контролю якості на всіх виробничих лініях, підвищуючи як ефективність, так і надійність.

Контроль якості є найважливішим аспектом виробництва, який безпосередньо впливає на надійність продукції та задоволення клієнтів і мультимодальний штучний інтелект кардинально змінює на краще цей процес, забезпечуючи перевірку в режимі реального часу, виявлення дефектів і адаптивний моніторинг.

Такі рішення, як Siemens Visual Inspection (рис 1.8), Bluebash, Cognex Multimodal або Landing AI, використовують комбінацію відеопотоку, даних із сенсорів глибини (LiDAR, ToF-камер). Системи зору на основі ШІ аналізують продукти безперервно, виявляючи дефекти поверхні, форми, вирівнювання чи маркування. Тепловізори фіксують відхилення у нагріванні або охолодженні, а вібраційні й акустичні датчики визначають аномалії в роботі обладнання. Об’єднуючи ці дані, мультимодальні моделі формують цілісне уявлення про якість і здатні не лише реагувати, а й **передбачати можливі дефекти**.

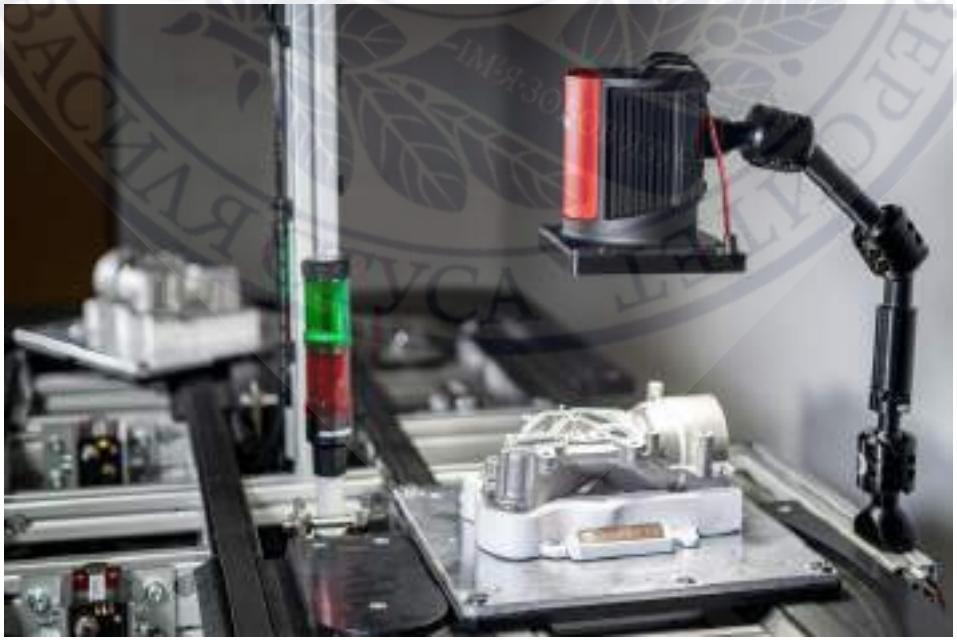

Рисунок 1.8 — Siemens Visual Inspection

Однією з ключових переваг є **перехресна перевірка між модальностями** — коли результати одного сенсора підтверджуються іншими. Наприклад, візуальне виявлення тріщини може бути перевірене вібраційними даними, а теплові аномалії — показниками датчиків. Це суттєво зменшує кількість помилкової продукції та підвищує точність рішень у процесах контролю якості.

Адаптивний моніторинг означає що система завдяки машинному навчанню з часом сама вивчає допустимі недоліки, визначає тенденції які можуть свідчити про поступове зниження якості та налаштовує параметри без втручання людини. Ця адаптивна здатність означає, що моделі можуть виявляти та виправляти проблеми **до того, як** вони вплинуть на цілі партії, що наперед значно зменшує ризики відходів та відкликання продукції [26].

1.3.3 Активне впровадження технологій мультимодального розпізнавання.

Мультимодальні технології розпізнавання вийшли далеко за межі академічних досліджень, активно трансформуючи ключові сектори економіки та суспільства.

У сфері **медицини** мультимодальність є критично важливою для підвищення точності діагностики та персоналізації лікування. Системи мультимодального розпізнавання інтегрують у медичні зображення (МРТ, КТ, УЗД, рентген) — для візуалізації та знаходження патологій та у лабораторні дані (результати аналізів крові, біопсії) — для біохімічної та молекулярної оцінки.

Наприклад DeepMind/Google Health AI — це автоматична ідентифікація та класифікація патологій.

Система розпізнає:

- Ураження/Новоутворення (пухлини, мікрокальцинати) на зображеннях.
- Патерни захворювань (наприклад, ознаки діабетичної ретинопатії на очному скані).
- Контекстуалізація знахідок шляхом зіставлення візуальних даних із ключовими термінами та клінічними описами з текстових звітів для підвищення точності діагнозу [27] .

Автономні транспортні засоби використовують мультимодальну інтеграцію для забезпечення безпечного та надійного руху.

Наприклад система розпізнавання об'єктів NVIDIA DRIVE AGX (або Hyperion) для автономоного транспотру — це високоточне 3D-розпізнавання та сегментація навколишнього середовища (рис 1.9).

Рисунок 1.9 — NVIDIA DRIVE AGX

Система розпізнає:

- Пішоходів та їхній рух (за допомогою камер та LiDAR)
- Транспортні засоби (відстань, швидкість, тип — через Радар та LiDAR).
- Дорожні знаки та розмітку (через камери). Інтеграція модальностей забезпечує надійність у складних погодних умовах (наприклад, LiDAR і Радар доповнюють камери при тумані) [28].

У літку 2025 року компанія NVIDIA здобула послідовну перемогу у конкурсі «End-to-End Autonomous Driving Grand Challenge» на конференції Computer Vision and Pattern Recognition (CVPR).

Як пише сама компанія, їм вдалося здобути перемогу серед інших учасників завдяки новітньому мультимодальному методу узагальненої оцінки траєкторій (GTRS) (рис 1.10), який генерує різноманітні траєкторії та поступово відфільтровує найкращі з них. Генерація безпечних та адаптивних траєкторій руху було доручено генерувати на основі даних кількох датчиків у напівреактивному моделюванні, де план автомобіля фіксований на початку, але фоновий рух динамічно змінюється.

GTRS пропонує комбінацію грубих наборів траєкторій, що охоплюють широкий спектр ситуацій, та дрібнозернистих траєкторій для критично важливих для безпеки ситуацій, створених за допомогою політики дифузії, обумовленої навколишнім середовищем. Потім GTRS використовує трансформаторний декодер, отриманий з метрик, залежних від сприйняття, зосереджуючись на безпеці, комфорті та дотриманні правил дорожнього руху. Цей декодер поступово фільтрує найперспективніші кандидати на траєкторії, фіксуючи тонкі, але критичні відмінності між подібними траєкторіями [29].

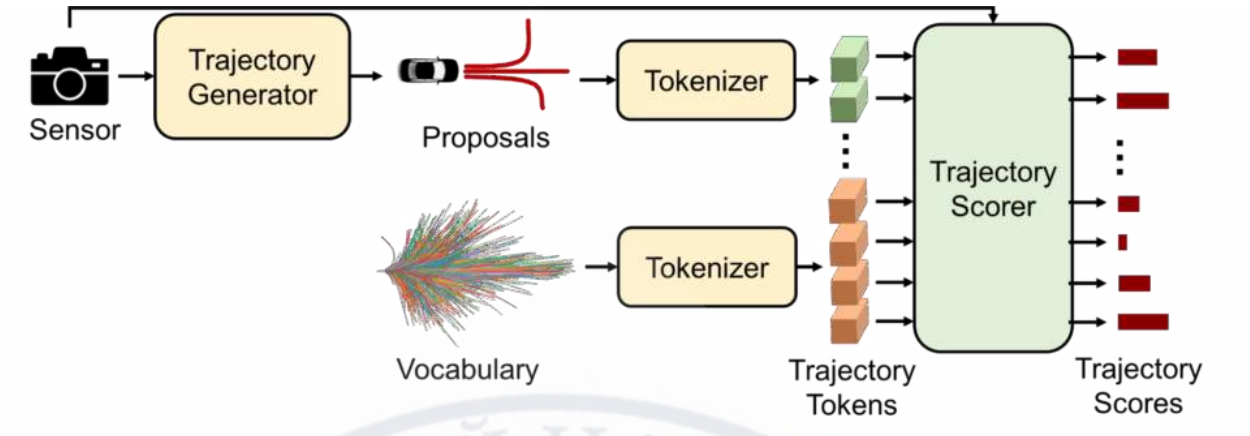

Рисунок 1.10 — Метод узагальненої оцінки траєкторій (GTRS)

Безпека та Спостереження

У системах безпеки мультимодальність використовується для всебічного моніторингу та швидкого реагування, основні завдання це:

- Розпізнавання осіб та об'єктів у складних умовах освітлення та високої зашумленості.
- Аналіз поведінки (виявлення нетипових дій, агресії, падінь) на основі інтеграції відео, звукових даних та теплових сигнатур. Наприклад у роботі “A Smart Security System Using Multimodal Features from Videos” описується як за допомогою аналізу мультимодальної біометрії, можна отримати систему безпеки, яка повністю невразлива до спуфінг-атак. [30].
- Багаторівневий контроль доступу (поєднання біометричних даних, голосу та зображення). Це підвищує надійність систем спостереження у громадських місцях та на критичних об'єктах інфраструктури.

Робототехніка та Взаємодія

Для робототехніки мультимодальне сприйняття є фундаментальним для навігації та взаємодії зі світом. Сенсорне сприйняття через зір (камери), дотик (тактильні сенсори для маніпуляторів), звук (визначення джерела та типу шуму) та пропріоцепцію (стан власних механізмів) грають вирішальну роль для розуміння середовища роботом. Інтеграція цих даних дозволяє роботам не лише будувати точні карти середовища, але й виконувати складні маніпуляційні завдання (наприклад, захоплення крихких об'єктів), де потрібна точна оцінка сили стискання на основі візуальної та тактильної інформації.

Наприклад існують Amazon Robotics (Picker Robots) (рис 1.11) або FANUC AI-Enabled Robots якаих основна функціональність це точне захоплення та сортування неідентичних об'єктів.

Рисунок 1.11 — Amazon Robotics (Picker Robot)

Вони здатні розпізнавати:

- Форму та геометрію об'єкта (визначення найкращої точки захоплення за допомогою 3D-зору).
- Матеріал, вагу та крихкість об'єкта (передбачення за допомогою зору, підтвердження та коригування сили стискання за допомогою тактильних датчиків). Це критично для роботи з об'єктами різної форми, які можуть бути ідентифіковані, але вимагають різного зусилля при маніпуляції [31].

Огляд сучасних мультимодальних інструментів демонструє, що розвиток технологій розпізнавання об’єктів невід’ємно пов’язаний із тенденцією до інтеграції різних типів даних у єдині інтелектуальні системи. Від промислових засобів контролю якості до великих мовно-візуальних моделей (LVLM), усі сучасні рішення ґрунтуються на спільному принципі — побудові єдиного векторного простору ознак, це дозволяє не лише підвищити точність і контекстність розпізнавання, але й розширити можливості взаємодії між людиною та машиною. Загалом, аналіз готових інструментів підтверджує, що мультимодальність стає визначальним напрямом розвитку систем розпізнавання об’єктів. Саме вона забезпечує основу для створення більш гнучких, контекстно орієнтованих і ефективних систем, що формують нове покоління інтелектуальних технологій штучного інтелекту.

1.4 Висновок до першого розділу

У першому розділі було проведено узагальнений аналіз сучасних підходів до мультимодального розпізнавання об’єктів. Розглянута основна теорія за темою, основні напрями розвитку мультимодальних систем розпізнавання об’єктів, розглянуті типи даних і як вони поєднуються. Огляд

сучасної літератури та готових інструментів показав, що мультимодальність стала стратегічним напрямом еволюції штучного інтелекту.

Отримані результати підкреслюють, що мультимодальне розпізнавання об’єктів є не просто продовженням розвитку комп’ютерного зору, а новою парадигмою штучного інтелекту — здатною поєднувати різні джерела інформації в єдину інтелектуальну систему. Це відкриває шлях до створення більш гнучких, надійних і контекстно орієнтованих технологій, які забезпечують якісно новий рівень взаємодії між людиною та цифровим середовищем.

РОЗДІЛ 2. ТЕХНІЧНІ ЗАВДАННЯ В МУЛЬТИМОДАЛЬНОМУ НАВЧАННІ

Мультимодальні системи стикаються з низкою базових технічних викликів, що виникають через різнорідність і різну інформативність модальностей, необхідність їх узгодження та побудови спільної семантичної структури. До таких викликів належать: гетерогенність типів даних і способів їхнього кодування; асинхронність або часткова невідповідність між модальностями; інформаційний дисбаланс, коли одна модальність є значно насиченішою або шумнішою за іншу; відсутність або неповнота даних; різні часові та просторові масштаби; а також потреба в оптимальному виборі рівня та механізму злиття ознак. Крім того, мультимодальні моделі повинні забезпечити побудову спільного простору представлень, у якому взаємопов’язані сигнали різних типів мають єдину геометрію та підпорядковуються спільним критеріям подібності. Реалізація такої системи вимагає ретельного проєктування енкодерів, механізмів ф’южну, ймовірнісних моделей спільного розподілу даних та методів оптимізації, здатних узгодити всі модальності в межах однієї архітектури. Тому далі в розділі буде розглянуто основні технічні завдання в мультимодальному навчанні для цього.

Ключовим технічним завданням у мультимодальному навчанні є вибір і побудова архітектури, здатної ефективно інтегрувати дані з різних модальностей, враховуючи їхню різнорідність, різну структуру та різний рівень інформаційної щільності. Незалежно від того, чи поєднуються зображення та текст, RGB і depth дані, мовні та аудіо сигнали — більшість сучасних систем дотримуються трирівневої парадигми, яка визначає загальну логіку конструювання моделей:

1) Представлення (Representation)

На першому етапі необхідно перетворити дані кожної модальності у компактний, інформативний та семантично насичений латентний

простір. Це здійснюється за допомогою модально-специфічних енкодерів: CNN або Vision Transformers для зображень, текстових трансформерів для природної мови, RNN/Transformer-архітектур для аудіосигналів, окремих 3D- або графових енкодерів для просторових даних (LiDAR, depth, IMU). Завдання цього етапу — зберегти максимально багатий зміст кожної модальності, але в узгодженій та формалізованій формі (векторах, послідовностях або тензорах ознак).

2) Злиття (Fusion)

Другий етап визначає, де і яким чином відбувається інтеграція закодованих представлень. Ф'южн може бути раннім, проміжним або пізнім; реалізовуватись через конкатенацію, гейти, механізми уваги, крос-модальні трансформери або багатопотокові структури. Це критичний момент, оскільки саме тут формується здатність моделі встановлювати відповідності між модальностями, комбінувати їхню комплементарність та пригнічувати шум або нерелевантні фрагменти. Від вибору схеми злиття залежить баланс між обчислювальною вартістю, інтерпретованістю та потужністю моделі.

3) Вирівнювання (Alignment)

Третій компонент полягає у конструюванні спільного простору вбудовування (shared embedding space), де елементи різних модальностей, що описують один і той самий об'єкт, розташовуються близько один до одного. Цей механізм критичний для пошуку відповідностей, кросмодального відображення, zero-shot задач і побудови універсальних мультимодальних моделей. На практиці вирівнювання досягається через спеціальні оптимізаційні функції (contrastive loss, triplet loss, InfoNCE), регуляризацію узгодженості або симетричне навчання енкодерів. Воно забезпечує модель здатністю «розуміти», що текстовий опис, RGB-зображення та depth-мапа можуть бути різними проекціями однієї сцени.

2.1 Представлення (Representation)

2.1.1 Формалізація модальностей як випадкових величин та просторів ознак.

У мультимодальних системах кожна модальність розглядається як випадкова величина, визначена на своєму вимірювальному просторі та породжуюча власний простір ознак. Така формалізація дозволяє математично описати, як різнорідні сигнали — зображення, текст, глибина, інфрачервоні дані чи аудіо — можуть бути узгоджені між собою у спільній моделі. Нехай X_i— модальність i-го типу (наприклад, RGB-зображення), тоді X_i— це випадкова величина зі значеннями у просторі $\mathcal{X}_i$, який може бути багатовимірним (простір тензорів пікселів), дискретним (послідовності токенів), або неперервним (інтенсивності глибинних сенсорів). Формально система працює з розподілом $p(X_1, X_2, \dots, X_n)$, що описує спільну статистику модальностей та їхню кореляцію. Докладніше цю концепцію описано в сучасних роботах про узгодження представлень мультимодальних систем [32].

Перехід від «сирих» модальностей до ознак здійснюється за допомогою відображення

$$f_i: \mathcal{X}_i \to \mathcal{Z}_i, \tag{2.1}$$

де $\mathcal{Z}_i$ — простір латентних ознак, характерний для кожної модальності.

Наприклад, для зображень f_iчасто задається згортковою мережею (ResNet, EfficientNet), а для тексту — трансформером (BERT, RoBERTa). Таким чином, кожна модальність має власне статистичне представлення,

але їх можна звести до узгоджених латентних просторів. Ідея полягає в тому, що хоч простори $\mathcal{X}_i$ фундаментально різні, простори $\mathcal{Z}_i$ вже можуть мати порівнювану топологію або навіть бути вкладені у спільний простір $\mathcal{Z}$. Саме так працює більшість моделей типу CLIP, де зображення та текст проектуються у спільний векторний простір, і подібність між ними вимірюється за допомогою косинусної міри [33].

Важливим математичним аспектом є опис спільної латентної структури. Нехай ϕ— функція узгодження, яка перетворює $\mathcal{Z}_i$ у єдиний простір ознак:

$$\phi_i(z_i) \in \mathcal{Z}, \tag{2.2}$$

де вже $\mathcal{Z}$ — спільний простір, у якому модальності можуть взаємодіяти.

Якщо представлення добре узгоджені, то вектори взаємопов'язаних даних (наприклад, текст «a dog» та зображення собаки) лежатимуть близько один до одного в $\mathcal{Z}$. Це дозволяє формально зрівнювати модальності через метрики відстаней: $d\big(\phi_1(z_1), \phi_2(z_2)\big)$. Формалізація модальностей у вигляді випадкових величин також дає можливість описувати взаємозалежності між ними через умовні розподіли. Наприклад, коли одна модальність слугує контекстом для іншої: $p(X_{image} \mid X_{text})$, що має ключове значення для моделей генерації (як у DALL-E чи Stable Diffusion) або для моделей з питання-відповіді, де текст визначає, яку частину зображення слід обробити [34]. Крім того, така формалізація дозволяє вводити припущення щодо незалежності або комплементарності модальностей, оцінювати ентропію $H(X_i)$, взаємну інформацію $I(X_i; X_j)$, та інші статистичні характеристики, які визначають користь або зайвість певних модальностей у системі. Висока взаємна інформація між модальностями означає, що моделі легше узгодити їх у спільний латентний простір, тоді як

низька — що вони надають унікальну, комплементарну інформацію, яку потрібно інтегрувати іншим чином. Ці концепції активно використовуються у сучасних роботах зі статистичного мультимодального навчання [35].

Підсумовуючи, мультимодальні дані описуються як набір випадкових величин із власними просторами ознак, які формують латентні представлення через функції перетворення. Узгодження цих просторів дозволяє моделі інтегрувати різні модальності у спільний семантичний простір, що є фундаментальною основою більшості сучасних мультимодальних архітектур.

2.1.2 Ймовірнісні моделі спільного розподілу $p(x_1, x_2, \dots, x_n)$.

Ймовірнісне моделювання мультимодальних даних ґрунтується на припущенні, що всі модальності пов'язані спільною прихованою структурою, яка визначає спільний розподіл $p(x_1, x_2, \dots, x_n)$. Кожна модальність розглядається як випадкова величина, а взаємозв'язки між модальностями визначаються тим, як інформація поширюється через цей спільний розподіл. На практиці це означає, що модель повинна вміти як описувати статистичну залежність між модальностями, так і працювати з умовними розподілами, наприклад $p(x_{image} \mid x_{text})$, що лежить в основі задач генерації, кросмодального пошуку або реконструкції пропущених модальностей. Подібний підхід докладно розглядається в літературі зі статистичного мультимодального навчання [36].

Спільний розподіл часто задається через приховані латентні змінні. Знову ж таки z— прихований фактор, який генерує всі модальності:

$$p(x_1, x_2, \dots, x_n, z) = p(z) \prod_{i=1}^{n} p\,(x_i \mid z\,), \tag{2.3}$$

де z — латентна змінна, що кодує спільну інформацію, притаманну всім модальностям;

$p(z)$— апріорний розподіл латентного простору (часто — нормальний $\mathcal{N}(0, I)$);

$p(x_i \mid z)$— генеративні розподіли для кожної модальності, що описують, як модальність x_iвиникає з латентного представлення.

У цьому випадку мультимодальна система намагається знайти таке латентне представлення z, яке пояснює варіації усіх модальностей одночасно. Цей підхід використовують варіаційні автоенкодери для мультимодальних даних, де оптимізація здійснюється через максимізацію евідансу нижньої межі (ELBO) [37].

У системах з двома модальностями (наприклад, текст–зображення) особливу роль відіграє взаємна інформація $I(X_1; X_2)$, яка визначає, наскільки сильно модальності корелюють між собою. Якщо взаємна інформація висока, то модель легше будує спільний простір ознак, і багато сучасних методів оптимізації, зокрема контрастивне навчання, прямо націлені на збільшення цієї інформаційної зв’язності. У роботах з представленнями CLIP та ALIGN ця ідея реалізована через максимізацію подібності між правильними парами та мінімізацію між випадковими.

Опис спільного розподілу також дозволяє працювати з неповними мультимодальними даними, що часто трапляється у реальних системах, де деякі модальності можуть бути відсутні. Спільна модель дозволяє використовувати апроксимацію умовного розподілу для відновлення пропущених даних:

$$p\left(x_j \mid x_i\right) = \int p\left(x_j \mid z\right) p\left(z \mid x_i\right) dz, \tag{2.4}$$

де x_i — спостережувана модальність

x_j — цільова відсутня модальність;

z — латентна змінна, спільний прихований простір модальностей;

$p(x_j \mid z)$ — генеративний розподіл модальності x_j;

$p(z \mid x_i)$ — апостеріорний розподіл;

$\int dz$ — інтегрування по всьому латентному простору;

Так реалізують **моделі «кросмодальної реконструкції»**, де одна модальність використовується для відновлення або прогнозування іншої — наприклад, генерація зображення з тексту або оцінка глибини з RGB-сцени [38].

У сучасних трансформерних системах спільний розподіл модальностей задається неявно через механізм attention, який моделює умовні залежності між елементами різних модальностей. У таких моделях, як LXMERT, мультимодальний attention фактично апроксимує умовні розподіли через матриці уваги, де сила взаємодії між елементами різних потоків інтерпретується як оцінка їхньої ймовірнісної залежності [39].

Окремий напрямок — це моделі, які безпосередньо оцінюють **спільний розподіл** без латентних змінних, наприклад енергетичні моделі. Вони визначають узгодженість модальностей через енергетичну функцію де нижча енергія відповідає більш реалістичним або статистично узгодженим конфігураціям модальностей. Енергетичні підходи активно використовуються у задачах кросмодального узгодження та генерації [40].

2.2 Злиття (Fusion)

Мультимодальний ф’южн дозволяє інтегрувати інформацію з різних джерел — візуальних, глибинних, мовних, аудіо, сенсорних тощо — створюючи більш інформативне та стабільне представлення.

Синергетичний ефект полягає в тому, що недоліки однієї модальності компенсуються сильними сторонами іншої, формуючи узгоджений латентний простір із кращими статистичними та семантичними властивостями.

Попри очевидні переваги, існує кілька принципово різних стратегій інтеграції модальностей: ранній (early fusion), пізній (late fusion), проміжний/гібридний (mid-level або hybrid fusion) та моделі на основі attention-орієнтованого узгодження. Кожен клас методів має власні переваги, обмеження та характерні випадки застосування, що потребує глибокої аналітики. Систематизація цих підходів важлива для коректного вибору архітектури мультимодальної системи у наукових і прикладних задачах, де необхідно враховувати структуру даних, вимоги до продуктивності, обчислювальні ресурси та очікувану точність.

У сучасних дослідженнях особливо помітний стрімкий розвиток трансформерних мультимодальних моделей, у яких інтеграція модальностей відбувається не через статичні операції (як-от конкатенація), а за рахунок динамічних механізмів cross-attention, co-attention та contextual alignment. Архітектури на кшталт ViLT, CLIP, Flamingo, BLIP-2, а також великі мультимодальні моделі GPT, Gemini, Claude показують, що багаторівневий attention-ф’южн дозволяє моделі формувати спільний контекст між модальностями, вирівнювати їх у єдиному семантичному просторі та підтримувати узгоджені висновки між зоровими та текстовими компонентами. Саме трансформери зумовили перехід від статичного ф’южну до контекстної, адаптивної та масштабованої інтеграції даних.

Класичні схеми мультимодального ф’южну — ранній та пізній — активно описані в літературі, особливо у роботах із комп’ютерного зору, аудіоаналізу та мультимодальної робототехніки. Ранній ф’южн добре працює в задачах, де модальності структурно узгоджені (наприклад, RGB та depth, синхронізовані з однієї камери), тоді як пізній ф’южн забезпечує більшу модульність, але може призводити до втрати міжмодальних

залежностей. Паралельно розвиваються проміжні та гібридні підходи, які поєднують кілька рівнів інтеграції, формуючи складні, ієрархічні стратегії багаторівневого ф’южну. Роботи останніх років демонструють, що єдиного універсального рішення не існує — оптимальна стратегія залежить від типу задачі, природи модальностей, масштабу даних та вимог до часу обробки [41].

2.2.1 Ранній ф’южн (Early Fusion)

Early fusion передбачає інтеграцію модальностей на рівні сирих даних або низькорівневих фіч. Наприклад:

- конкатенація каналів: RGB + D 4-канальне зображення,
- об’єднання точкових хмар із інтенсивністю, кольором чи нормалями,
- синхронізоване подання аудіо та спектрограм,
- формування спільного сенсорного простору.

Такий підхід створює єдине латентне представлення, яке обробляється однією нейронною мережею. Це дозволяє:

- зберегти максимальний обсяг крос-модальної інформації,
- навчати перші шари моделі виявляти складні міжканальні залежності,
- створити модель, яка "бачить" дані цілісно, а не окремими потоками.

Основні переваги:

- максимальне збереження кореляцій між модальностями;
- високий потенціал для задач локалізації та детекції;
- природне формування спільного простору ознак на ранніх етапах.

Недоліки:

- високі вимоги до точної синхронізації, калібрування та нормалізації модальностей;
- чутливість до шуму, адже помилка в одній модальності впливає на спільне представлення;
- ризик “домінування” однієї модальності, якщо вона статистично сильніша.

Ранній ф’южн найчастіше використовується в задачах, де модальності структурно узгоджені в просторі (RGB-D, LiDAR + камери, мультиспектральні зображення).

2.2.2 Пізній ф’южн (Late Fusion)

Late fusion інтегрує дані для високорівневих рішень окремих одномодальних моделей. Зазвичай це:

- усереднення ймовірностей,
- вагове агрегування,
- voting-методи,
- метакласифікатор поверх модальних прогнозів.

Переваги:

- модульність системи, що дозволяє незалежно тренувати й замінювати модальні моделі;
- можливість простого масштабування — додавання нової модальності не змінює існуючу архітектуру;
- зручна інтеграція pre-trained моделей без потреби їх глибокої модифікації.

Недоліки:

- модель не бачить спільний простір ознак — міжмодальні залежності втрачаються;
- слабка ефективність у задачах, де важливе точне взаємне вирівнювання (наприклад, детекція дрібних об’єктів).

Late fusion переважно застосовується у різнорідних, неструктурованих або асинхронних даних, а також там, де важлива надійність окремих модулів.

2.2.3 Гібридний та проміжний ф’южн (Hybrid / Mid-Level Fusion)

Гібридні методи виконують інтеграцію на кількох рівнях — низькому, середньому та високому:

- об’єднання фіч-мапів на певних шарах CNN,
- інтеграція поетапно на різних масштабах,
- спільне навчання латентних векторів та пізнє агрегування прогнозів,
- багаторівнева інтеграція в трансформерах.

Цей підхід дає:

- баланс між точністю раннього ф’южну та гнучкістю пізнього;
- здатність моделі виявляти складні кореляції, не втрачаючи модульності;
- можливість тонкого налаштування взаємодії модальностей.

Однак гібридні системи:

- складніші в оптимізації,
- потребують більше пам’яті й обчислень,
- можуть бути чутливими до нерівномірної інформативності модальностей.

2.2.4 Attention-орієнтований ф’южн та трансформерні методи

Attention-ф’южн інтегрує дані адаптивно: модель сама визначає, коли і яку модальність використовувати, підсилюючи релевантні ознаки та приглушуючи шумові. Основою є механізми:

- cross-attention,
- co-attention,
- multi-head fusion,
- contextual alignment,
- token-level multimodal binding.

Переваги:

- здатність моделювати складні, контекстно залежні кореляції;
- масштабування на великі модальні простори (тисячі токенів і зображень);
- автоматичне вирівнювання простору ознак без жорсткої прив’язки до геометрії;

- state-of-the-art результати у візуально-мовних моделях, reasoning та 3D-аналізі.

Недоліки:

- дуже велика обчислювальна вартість,
- потреба у великих датасетах,
- складність інтерпретації міжмодальних залежностей.

Attention-ф’южн став стандартом у LLM з візуальними можливостями, мультимодальних агентів та сучасних моделей розуміння сцени [42].

Отже, злиття модальностей дозволяє інтегрувати різні джерела інформації для досягнення точніших результатів. Кожна стратегія злиття — рання, пізня, гібридна та attention-орієнтована — має свої переваги та обмеження. Ранній ф’южн зберігає кореляції, але вимагає точної синхронізації, пізній — більшу гнучкість, але може втрачати міжмодальні зв’язки. Гібридні підходи поєднують переваги обох методів, а attention-ф’южн забезпечує адаптивне злиття. Вибір стратегії залежить від задачі, типу даних і вимог до ефективності.

2.3 Вирівнювання модальностей (Alignment)

Фундамент мультимодального навчання полягає у створенні спільного або узгодженого простору представлень, у якому різні модальності (наприклад, текст і зображення) можуть бути співставлені, порівняні або інтегровані. Для цього застосовуються дві ключові групи математичних інструментів: **метрики подібності**, що визначають ступінь близькості між векторами представлень, та **оптимізаційні функції**, які спрямовують модель на формування структурованого, узгодженого латентного простору. Ці два аспекти завжди працюють разом: метрики визначають, що означає «бути схожими», а функції втрат — як саме модель повинна навчатися такої схожості.

Однією з базових метрик є **косинусна подібність (cosine similarity)**, яка вимірює кут між векторами в багатовимірному просторі, що робить її особливо ефективною для порівняння представлень різних джерел. Такий підхід є основою великої кількості сучасних мультимодальних моделей, зокрема CLIP, де косинусна подібність визначає відповідність тексту й зображення у спільному просторі ознак [43].

Поряд із цим широко використовується **дивергентність Кульбака–Лейблера (KL-divergence)**, яка слугує мірою відмінності між ймовірнісними розподілами. У мультимодальних моделях вона застосовується для узгодження латентних просторів різних модальностей, особливо в архітектурах VAE, де необхідно забезпечити спільність статистичних властивостей у прихованих змінних.

Метричні поняття доповнюються оптимізаційними функціями, які визначають механізм навчання узгоджених представлень. Одним із ключових підходів є **contrastive learning**, метою якого є зближення позитивних кросмодальних пар та відштовхування негативних. Його сучасні формулювання (наприклад, InfoNCE) відіграють центральну роль у тренуванні моделей CLIP, ALIGN та інших систем вирівнювання текст–зображення [44].

Важливим інструментом також є **triplet-loss**, який вимагає, щоб представлення «anchor» було ближчим до правильного «positive» прикладу, ніж до «negative», з певним відступом (margin). У мультимодальних сценаріях це дозволяє будувати чіткі відстані між семантичними та несемантичними парами, що зміцнює структурованість мультимодального простору [45].

Ще одним підходом є використання крос-ентропійної функції втрат між модальностями, яка використовується в задачах, де одна модальність передбачає іншу (наприклад, текст → категорія зображення). Вона дозволяє оптимізувати модель у напрямку кросмодального передбачення, забезпечуючи узгодження через класифікаційні структури.

Не менш важливими є принципи alignment/uniformity, які були запропоновані в сучасній теорії контрастивного навчання. Alignment це частина де елементи різних модальностей повинні бути близькими. Uniformity це де всі представлення мають рівномірно заповнювати простір, щоб уникнути колапсу моделі. У мультимодальних системах ці дві вимоги балансуються: модель одночасно навчається зближувати коректні пари та підтримувати розподіленість ознак на глобальному рівні. Це формує добре структурований, семантично значущий простір представлень.

Таким чином, метрики подібності та оптимізаційні функції становлять єдиний математичний каркас, який визначає, як мультимодальні представлення взаємодіють, узгоджуються та набувають семантичного значення. Метрики формалізують поняття близькості, а функції втрат керують процесом навчання, примушуючи простір представлень бути одночасно узгодженим на локальному рівні (alignment) і стійким на глобальному (uniformity) [46]. Це робить можливим створення мультимодальних моделей, здатних глибоко розуміти та поєднувати інформацію з різноманітних джерел.

2.4 Висновок до другого розділу

У цьому розділі було сформульовано фундаментальні технічні завдання мультимодального навчання, що визначають архітектуру, поведінку та обмеження сучасних систем. Розглянуто три ключові компоненти — представлення, злиття та вирівнювання модальностей — які разом формують універсальну парадигму побудови мультимодальних моделей.

РОЗДІЛ 3. ВИБІР ІНСТРУМЕНТІВ АРХІТЕКТУР ТА МЕТОДІВ ДЛЯ РЕАЛІЗАЦІЇ

Вибір мультимодальної архітектури, методів об’єднання модальностей та наборів параметрів навчання є ключовими елементами даного розділу.

На початковому етапі необхідно визначити базову модель для подальшого розширення під мультимодальні вхідні дані. Обрана архітектура має відповідати ряду критеріїв: наявність переднавчених ваг, сучасність, стабільність, можливість модифікації початкових шарів та зрозумілий експериментальний протокол. Саме тому наступний підрозділ присвячено обґрунтуванню вибору базової архітектури, яка слугуватиме основою для розробки мультимодального рішення

3.1 Вибір базової архітектури

Вибір базової моделі є ключовим етапом у побудові мультимодальної системи детекції, оскільки саме вона визначає як початкову якість роботи, так і потенціал подальшої інтеграції додаткових модальностей. Для мультимодальної детекції, що комбінує RGB-зображення та карти глибини, важливо, щоб вибрана архітектура мала такі властивості:

- наявність переднавчених ваг на великому датасеті
- висока стабільність та популярність, що забезпечує надійність та відтворюваність результатів;
- гнучкість структури, зокрема можливість модифікації вхідного шару та backbone для інтеграції додаткового каналу глибини;
- достатня сучасність, щоб модель відповідала сучасному рівню досліджень у сфері детекції;

- підтримка від спільноти та розробників

Враховуючи ці вимоги, у цій роботі як базову модель обрано **YOLOv5** — одну з найбільш поширених і стабільних моделей детекції об’єктів останніх років. Її популярність спричинена збалансованим поєднанням точності, швидкості та простоти адаптації. Важливо, що YOLOv5 має повністю відкритий вихідний код, офіційні переднавчені ваги та гнучку архітектуру, що дозволяє безпосередньо розширяти вхідний шар, додаючи новий мультимодальний канал, що дає можливість виконувати дослідження.

3.1.1 Архітектура YOLOv5 (You Only Look Once)

Основною задачею **YOLOv5** є виявлення об'єктів, що передбачає виділення суттєвих ознак з вхідних зображень. Ці ознаки потім обробляються прогностичною моделлю для локалізації та класифікації об'єктів на зображенні. Архітектура YOLO ввела підхід **end-to-end** (від початку до кінця), що дозволяє навчати модель на єдиному диференційованому контексті, об'єднуючи завдання регресії обмежувальних коробок та класифікації об'єктів у єдину нейронну мережу [47].

Фундаментально, мережа **YOLOv5** складається з трьох основних компонентів (рис 3.1):

1. **Backbone** (основна частина) — згорткова нейронна мережа, яка відповідає за кодування інформації зображення у вигляді **feature maps** (карти ознак) на різних масштабах. Ці карти ознак містять інформацію про низькорівневі деталі зображення, такі як краї, текстури та інші важливі ознаки. використовується варіант CSPDarknet53, який є модифікацією популярної Darknet53,

адаптованої для швидшого навчання і обробки. CSPDarknet53 — це згорткова нейронна мережа, яка включає в себе CSP (Cross-Stage Partial) модуль для зменшення обчислювальних витрат і підвищення ефективності. CSP модуль дозволяє уникнути надмірної кількості обчислень і редундантної інформації, ефективно працюючи з градієнтами під час навчання. Backbone використовує **Batch Normalization** для нормалізації активацій та зменшення внутрішнього зміщення під час навчання, що покращує стабільність і швидкість навчання.

2. **Neck** (шиї) — це серія шарів, призначених для інтеграції та уточнення представлених ознак, отриманих з backbone. Neck дозволяє з'єднати інформацію з різних рівнів і розрізів карти ознак, що підвищує ефективність подальшої обробки та покращує точність. В цій частині використовується модифікація **PANet**, яка є основною архітектурою neck. PANet дозволяє ефективно агрегувати ознаки на різних масштабах і допомагає вирішити проблему локалізації об'єктів на великих та малих масштабах. PANet використовує механізм **об'єднання різноманітних шляхів (path aggregation)** для покращення передачі ознак між шарами та досягнення більш точної локалізації.

3. **Head/dense prediction** (голова) — на цьому етапі мережа генерує прогнози для bounding boxes та класів об'єктів на основі оброблених ознак. Це кінцева стадія, яка визначає, де саме на зображенні знаходяться об'єкти і до якої категорії вони належать. Архітектура YOLOv5 передбачає координати обмежувальної рамки як відхилення від певних заздалегідь визначених розмірів, які використовуються як орієнтири. Ці розміри рамок важливі для початку прогнозування і можуть значно впливати на ефективність роботи моделі.

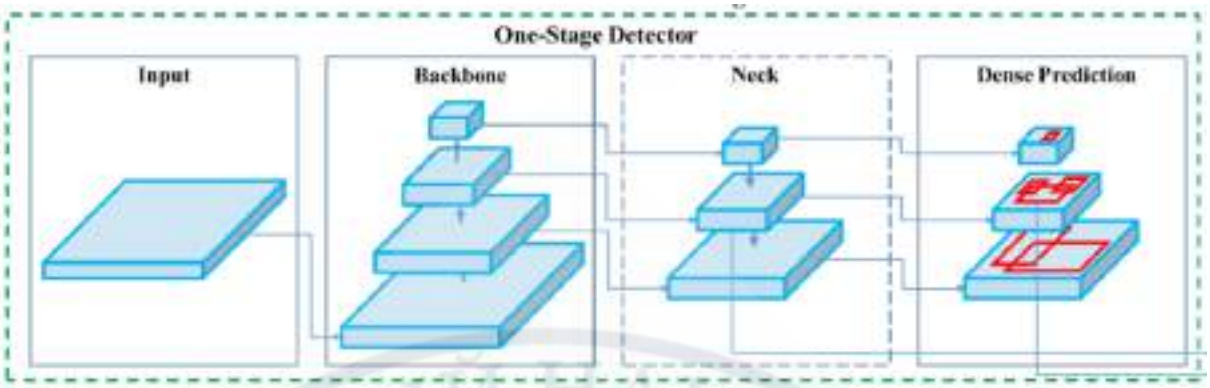

Рисунок 3.1 — Архітектура YOLOv5

YOLOv5 є значним кроком уперед, оскільки переносить архітектуру YOLO з фреймворку Darknet на PyTorch. Фреймворк Darknet, що в основному реалізований на мові C, надає дослідникам детальний контроль над операціями мережі. Хоча цей рівень контролю є корисним для експериментів, він часто ускладнює швидку інтеграцію нових досліджень через необхідність виконання кастомних обчислень градієнтів для кожної нової реалізації. Перехід на фреймворк PyTorch значно допоможе у практичній частині.

Функція втрат YOLOv5 є складною і складається з трьох компонентів: Binary Cross-Entropy (BCE) для прогнозування класу та об'єктності, а також Complete Intersection over Union (CIoU) для локалізації. Загальна втрата обчислюється як зважена сума цих окремих втрат:

$$\text{Loss} = \lambda_1 \cdot L_{\text{cls}} + \lambda_2 \cdot L_{\text{obj}} + \lambda_3 \cdot L_{\text{loc}}, \tag{3.1}$$

де L_{cls}, L_{obj}і L_{loc}— це відповідно втрата Binary Cross-Entropy для прогнозування класу, втрата Binary Cross-Entropy для об'єктності та втрата Complete Intersection over Union для локалізації.

Коефіцієнти λ_1, λ_2та λ_3 — це гіперпараметри, які балансують внесок кожної компоненти втрат в загальний процес оптимізації [48].

Крім того, YOLOv5 демонструє високу ефективність та можливість зміни архітектури моделі для перенавчання моделі на невеликих і середніх датасетах, що є критично важливим для цього дослідження. Тому для створення мультимодальності будемо використовувати **підхід fine-tuning.**

3.1.2 Fine-tuning

Fine-tuning (до-налаштування) — це метод адаптації переднавченої нейронної мережі до нової задачі або нового датасету шляхом подальшого навчання моделі на обмеженій кількості нових даних (рис 3.2). У науковій літературі fine-tuning належить до підкласу transfer learning, коли знання, отримані на великому джерельному датасеті, передаються на іншу цільову задачу [49].

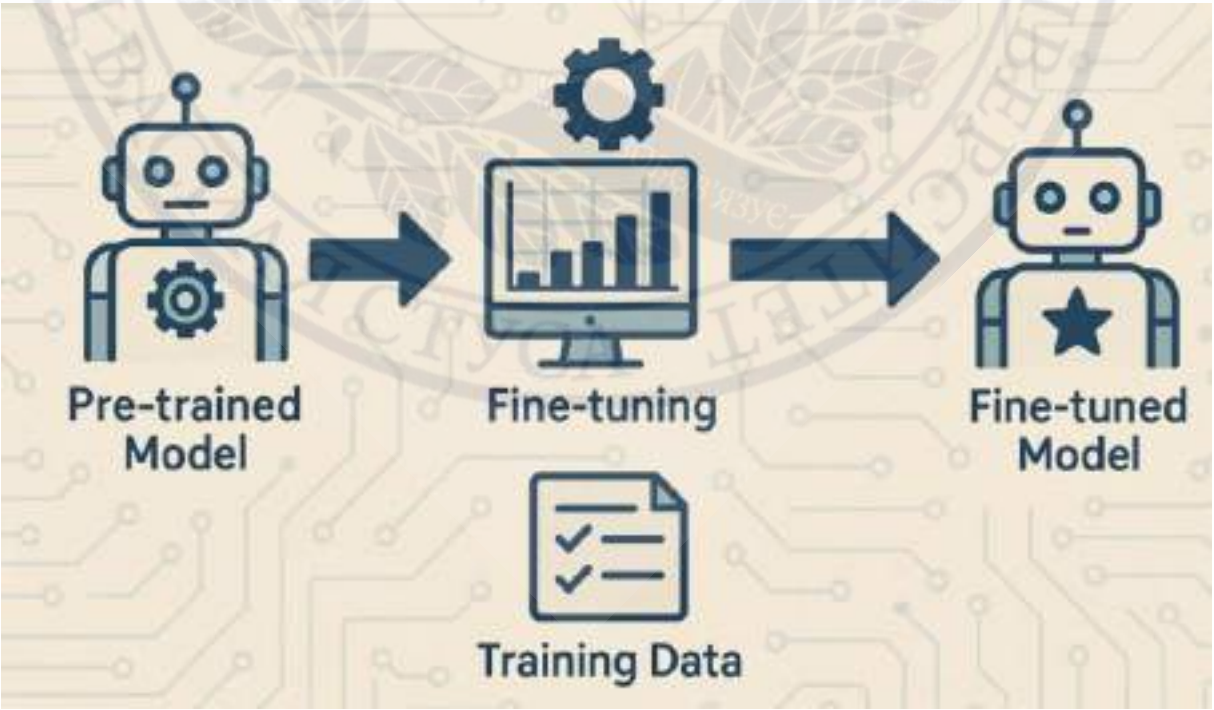

Рисунок 3.2 — Схема до-налаштування моделі

У роботах з transfer learning показано основну характеристику глибоких згорткових моделей, що вони формують універсальні низькорівневі ознаки (edges, textures, corners), які можна повторно використовувати без повного перенавчання моделі [50]. Це дуже важливо, адже глибокі мережі часто мають мільйони параметрів, і їх навчання на великій кількості даних може бути надзвичайно витратним за часом та обчислювальними ресурсами.

В рамках **transfer learning**, існує кілька типів підходів, залежно від того, яку частину моделі адаптують:

1. **Повне перенавчання (full fine-tuning).**: У цьому випадку, після адаптації моделі до нової задачі, модифікуються всі шари мережі (як верхні, так і нижні). Це дозволяє моделі повністю "перенавчитися" на новому датасеті, зберігаючи при цьому базові ознаки, отримані на попередньому наборі даних.
2. **Перенавчання лише верхніх шарів**: В більшості випадків, особливо в задачах з обмеженими даними, зазвичай здійснюється fine-tuning лише на останніх шарах нейронної мережі. Це обумовлено тим, що верхні шари моделі часто відповідають за загальні ознаки, такі як текстури, краї та кольори, які є універсальними для різних типів даних. Останні шари, в свою чергу, спеціалізуються на більш складних абстракціях, притаманних конкретній задачі.
3. **Розморожування лише кількох останніх шарів або часткове перенавчання**: Існує також варіант, коли заморожуються всі шари, окрім останніх кількох. Такий підхід використовується, коли є потреба зберегти більшу частину знань моделі яка відповідна за загальні ознаки але і одночасно адаптувати мережу до нової задачі.

Сучасні параметр-ефективні методи адаптації (PEFT)

Для зменшення обчислювальних витрат та підвищення стабільності навчання були розроблені методи, що дозволяють адаптувати модель *без зміни її основних ваг*:

- **Additive fine-tuning.** До моделі додаються додаткові параметри або адаптаційні блоки, тоді як всі базові ваги залишаються замороженими. Це дає змогу заощадити ресурси, але забезпечити додаткову гнучкість.
- **Reparameterization (LoRA).** Оновлення ваги подається не напряму, а через низькорангові матриці, які накладаються на основні параметри вже під час інференсу. Базові ваги не змінюються, а адаптація виконується через незалежні компактні матриці A та B. Це дає змогу створювати кілька «версій» адаптації для різних задач.
- **Partial fine-tuning у контексті lightweight-методів.** Інколи цей термін використовується для ситуацій, коли тренуються лише зовнішні (високорівневі) блоки, а вся базова архітектура залишається незмінною. Такий підхід по суті є feature extraction.

Ці методи є важливими для дуже великих моделей або коли є обмеження на обсяг тренувальних даних [51].

Переваги fine-tuning

1. **Швидкість навчання**: Оскільки більша частина навчання вже проведена на великому наборі даних, fine-tuning дозволяє значно зменшити час, необхідний для адаптації моделі до нових умов. Навчання «з нуля» потребує більше часу, оскільки модель повинна вивчити всі ознаки та абстракції самостійно.
2. **Зменшення обсягу даних**: Для до-налаштування моделі необхідно набагато менше нових даних. Це особливо корисно в умовах обмеженого доступу до даних, коли неможливо зібрати великий та різноманітний датасет для навчання з нуля.

3. **Покращення стабільності**: Fine-tuning дозволяє зберегти стабільність навчання, оскільки модель вже має базову здатність до узагальнення, отриману на попередньому етапі. Це означає, що в процесі fine-tuning модель менш схильна до переобучення на обмежених даних.
4. **Поліпшення результатів на складних задачах**: Fine-tuning дозволяє досягти високих результатів навіть на складних задачах, де специфічні дані можуть мати різні варіації або обмежену кількість прикладів. Це є важливим у таких сферах, як медична візуалізація, де кожен набір зображень є унікальним.

Отже, метод fine-tuning дозволяє значно скоротити час тренування, зменшити обсяг необхідних даних та підвищити стабільність навчання порівняно з тренуванням «з нуля». Він є оптимальним підходом для адаптації переднавченої моделі YOLOv5 до мультимодального вхідного формату RGBD, оскільки дає змогу зберегти корисні ознаки, вивчені на COCO2017, плавно інтегрувати depth-канал, мінімізувати обчислювальні витрати та досягти кращої якості детекції без необхідності у великому власному датасеті.

3.2 Використані датасети в роботі

Одним з ключових компонентів у побудові системи детекції об’єктів є вибір репрезентативного та достатньо масштабного датасету. Важливість вибору правильного датасету полягає в тому, що він визначає здатність моделі до узагальнення на нові, раніше не бачені дані. Якщо датасет не є репрезентативним або недостатньо великим, модель може перенавчатися на специфічних прикладах або не здатна адекватно працювати в реальних умовах. Наприклад, якщо датасет містить лише об'єкти, зняті при одному

типі освітлення або в одному середовищі, модель може погано працювати в інших умовах, де освітлення або фони змінюються.

Також важливою є різноманітність об'єктів на зображеннях: датасет має включати різні варіації кожного об'єкта, наприклад, в різних ракурсах, з різними рівнями перекриття або в присутності інших об'єктів, щоб навчити модель правильно ідентифікувати та локалізувати об'єкти в складних умовах. Різноманітність фонових умов, видів і стилів зображень робить модель більш універсальною, а також підвищує її здатність до адаптації в різних ситуаціях.

Датасет COCO має в собі усі показники підходящого датасету, тому саме на такому датасеті була переднавчена модель YOLOv5 [52].

Датасет COCO2017 (Common Objects in Context) містить різноманітні зображення об’єктів у реальних сценах, що охоплюють широкий спектр умов зйомки, масштабів, ракурсів та ступенів оклюзії. Його основною метою є забезпечення складного та реалістичного середовища для навчання моделей детекції, сегментації та позначення ключових точок (рис 3.3).

Структура COCO2017 включає два основні піднабори:

- train2017 — 118 000 зображень,
- val2017 — 5 000 зображень.

Зображення мають детальну та стандартизовану розмітку у форматі JSON – це позначення 80 класів об’єктів, включаючи людей, транспорт, тварин, побутові предмети тощо та координати класів об'єктів (bounding boxes) [53].

Рисунок 3.3 — Прикдал датасету COCO

Оскільки базовий датасет COCO2017 містить лише RGB-зображення та не включає каналу глибини, для реалізації мультимодального підходу необхідно знайти або сформувати **COCO-Depth** набір даних. Для виконання практичної частини роботу було обрано саме глибину згенеровану за допомогою сучасної нейромережевої моделі **MiDaS (**monocular depth estimators) — одного з найточніших монокулярних оцінювачів глибини.

Модель MiDaS дозволяє обчислювати відносну глибину сцени за одним RGB-зображенням (рис 3.4). На відміну від стерео- або LiDAR-систем, MiDaS не потребує апаратної глибини, що робить її універсальним

інструментом для великих датасетів, де вже неможливо зібрати реальні depth-дані [54].

Модель навчається на величезних базах даних, що містять зображення різних сцен з відомими значеннями глибини, і намагається зрозуміти, як глибина об'єктів залежить від їх розміру, текстури, освітлення і просторової орієнтації.

MiDaS використовує сучасні архітектури глибоких нейронних мереж для виділення ознак зображення. Потім мережа передає ці ознаки через кілька шарів для отримання фінальної оцінки глибини. Використовується також техніка **"fully convolutional"** для забезпечення високої точності на різних масштабах зображення.

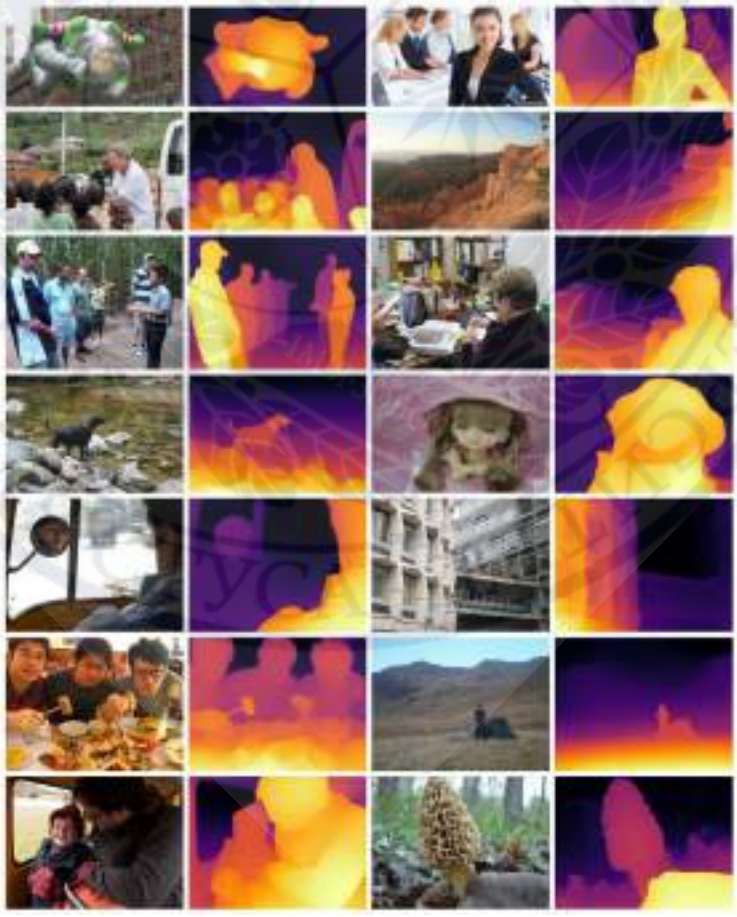

Рис 3.4 – приклад роботи моделі MiDaS

Синтезований depth-канал є градаційним зображенням, у якому світлі області відповідають ближчим об'єктам, а темні — далеким. Така репрезентація добре підходить для інтеграції разом із RGB як четвертий канал у моделі типу YOLO.

Для отримання модальності "глибина" всі зображення COCO2017 були послідовно пропущені через модель MiDaS. У результаті сформовано два піднабори depth-зображень, що повністю відповідають структурі оригінального COCO:

- train2017_depth — 118 000 depth-карт,
- val2017_depth — 5 000 depth-карт.

Кожне depth-зображення в датасеті має **ідентичну назву,** розмір й відповідне RGB-зображення. Загалом ідентичні назви в додатковому датасеті для ще одного каналу можуть значно полегшити процес fine-tuning для вже натренованої моделі через те, що:

- ідентичність назв каналів між основним та додатковим датасетами дозволяє зберегти узгодженість між навчальними даними. Це означає, що модель, яка вже має певне розуміння категорій та ознак, може легко застосувати набуті знання до нових даних, що містять ті ж самі назви. У цьому випадку модель не потребує повного перенавчання на новий канал, оскільки вона вже має попередній досвід роботи з подібними категоріями.
- збереження таких назв забезпечує збереження контексту. Якщо модель вже була натренована на даних певного каналу, вона може ефективно використовувати знайомі ознаки при обробці нових даних, що дозволяє **покращити точність** та **ефективність** на новому каналі без втрати знань про попередні канали.
- подібні назви сприяють покращенню **узагальнення** моделі. Оскільки модель може виявляти спільні патерни між каналами з ідентичними

назвами, вона здатна швидше адаптуватися до нових умов, навіть якщо канали мають незначні відмінності у змісті. Це дозволяє моделі краще справлятися з варіативністю даних і застосовувати набуті знання для прогнозування.

Глибинний датасет COCO-Depth не містить власних анотацій, оскільки його єдина роль — бути додатковою модальністю до існуючих RGB-даних COCO [55]. Тому Анотації беруться повністю з COCO2017. Depth-карти синхронізовані з RGB-файлами за назвами файлів. З врахуванням цих переваг було обрано і використано саме цей датасет.

Таким чином, правильний вибір датасету — це ключ до успіху системи комп'ютерного зору. Він дозволяє не лише ефективно навчити модель, але й гарантувати її здатність до роботи в реальних, змінних умовах.

3.3 Вибір бібліотек та інструментів реалізації

Під час розробки мультимодальної системи розпізнавання об'єктів на основі RGB та глибинних зображень (RGB-D) було обрано низку програмних інструментів, бібліотек та фреймворків, що забезпечують ефективне навчання, обробку даних та оцінку результатів. Добір інструментів визначався вимогами до продуктивності, зручності інтеграції та масштабованості експериментів.

Система створювалась з використанням мови програмування Python. Основним фреймворком для побудови та навчання нейронних мереж у роботі є PyTorch, який забезпечує:

- динамічну обчислювальну графіку (зручність експериментів і модифікацій);
- широкі можливості кастомізації архітектури YOLOv5;
- підтримку GPU через CUDA;

велику екосистему доповнень та готових компонентів (TorchVision, timm). TorchVision в свою чергу забезпечив базову обробку зображень, завантаження та попередню обробку датасетів.

YOLOv5 офіційно реалізований у PyTorch, що робить його природним вибором для проведення fine-tuning та експериментів із раннім злиттям RGB та Depth.

Також використовувались такі бібліотеки для для аналізу та візуалізації:

- Бібліотека OpenCV застосовувалась для читання та конвертації зображень,
- NumPy використовувався для роботи з даними та оцінки якості даних;

Ultralytics це компанія, яка розробляє інструменти та бібліотеки для комп'ютерного зору, зокрема для задач детекції об'єктів, розпізнавання зображень та перетренування глибоких нейронних мереж. Компанія також має власну бібліотеку на Python [56].

В роботі використано офіційну реалізацію Ultralytics YOLOv5 з відкритим вихідним кодом:

- забезпечує гнучку конфігурацію архітектури,
- дозволяє модифікувати вхідні канали backbone,
- має модульну структуру для адаптації процесу тренування.

Для системи створення було обране робоче середовище **Conda**, де було створено окреме conda-середовище для ізольованого встановлення залежностей, контролю версій бібліотек та стабільності у виконанні.

Комплекс цих інструментів забезпечив ефективну підготовку даних, гнучку модифікацію архітектури, стабільне навчання та ґрунтовну оцінку результатів.

3.4 Висновок до третього розділу

У цьому розділі було обґрунтовано вибір архітектурних рішень, методів обробки даних та програмних інструментів, що стали основою для реалізації мультимодальної системи детектування об’єктів. На першому етапі було обрано базову модель YOLOv5. Детально проаналізовано її структуру (backbone–neck–head) та механізми роботи, а також визначено стратегію тонкого донавчання (fine-tuning).

На другому етапі було розглянуто які датасети будуть використовуватись в роботі,а також загальний опис, їх структуру, методи створення датасету.

На третьому було здійснено обґрунтований вибір інструментів, фреймворків та бібліотек.

РОЗДІЛ 4. ПРАКТИЧНА РЕАЛІЗАЦІЯ МУЛЬТИМОДАЛЬНОЇ МОДЕЛІ РОЗПІЗНАВАННЯ ОБ'ЄКТІВ

Практичний розділ магістерської роботи спрямований на проєктування, реалізацію та експериментальне дослідження мультимодальної системи розпізнавання об'єктів, що поєднує вхідні дані RGB та оцінку глибини. Основна мета експериментів — продемонструвати, що додавання карти глибини, отриманої за допомогою алгоритму MiDaS, може покращити роботу сучасних моделей детекції, а також дослідити ефективність різних стратегій мультимодальної інтеграції, зокрема підходу early fusion (рис 4.1).

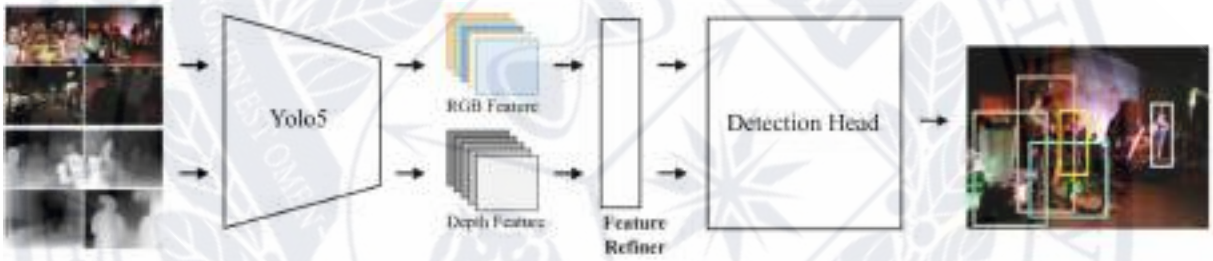

Рисунок 4.1 — Схема мультимодальної моделі

4.1 Створення моделі

Для налаштування програмного оточення для роботи з **YOLOv5**, важливо створити ізольоване середовище, що дозволяє уникнути конфліктів між різними бібліотеками та гарантує стабільність роботи проекту. Встановлення залежностей та правильні версії бібліотек є критично важливими для коректної роботи моделі. Було створено нове **Conda**-середовище, яке дозволить ізолювати проект від інших бібліотек.

Конда (Conda) — це одна із самих популярних систем управління середовищами та пакетами, що дозволяє ізолювати проекти та керувати залежностями.

Помісти папки train_depth та val_depth разом з оригінальним COCO images, та подальша нормалізація зображень.

Нормалізація

```
def normalize_depth_and_save(depth_array, out_path):
    p1, p99 = np.percentile(depth_array, (1, 99))
    depth_clipped = np.clip(depth_array, p1, p99)
    depth_norm = (depth_clipped - p1) / (p99 - p1 + 1e-6)  # 0..1
    depth_uint16 = (depth_norm * 65535).astype('uint16')
    Image.fromarray(depth_uint16).save(out_path)
```

В подальшому необхідно модифікувати завантаження даних у модель, тобто додати додатковий канал глибини. YOLOv5 зчитує завантажені дані з datasets.py (файл utils/datasets.py). Потрібно змінити блок, де зчитується звичайне зображення і туди додати читання depth з конкатенацію.

```
def load_rgbd(rgb_path, depth_path):
    rgb = cv2.imread(rgb_path)
    depth = cv2.imread(depth_path, cv2.IMREAD_UNCHANGED)
    if depth.dtype == np.uint16:
        depth = depth.astype('float32') / 65535.0
    else:
        depth = depth.astype('float32')
        depth = (depth - depth.min()) / (depth.max()-depth.min()+1e-6)
    depth = cv2.resize(depth, (rgb.shape[1], rgb.shape[0]))  # match size
    depth = depth[:, :, None]
    rgbd = np.concatenate([rgb, depth*255.0], axis=2
    return rgbd.astype('uint8')
```

Далі вже необхідно змінити саму архітектуру YOLOv5, щоб приймати 4 канали. Одночасно із цим було змінено конфігураційний файл для четвертого каналу.

```
import torch

# load pretrained yolov5 weights
ckpt = torch.load('yolov5s.pt', map_location='cpu')
model_state = ckpt['model'].float().state_dict()

from models.yolov5s import Model
model = Model('models/yolov5s.yaml')  # ch=4 в yaml
ms = model.state_dict()

for k,v in model_state.items():
    if v.ndim==4 and v.shape[1]==3:
        first_key = k
        first_weight = v
        break

new_weight = ms[first_key]
with torch.no_grad():
    new_weight[:, :3, :, :] = first_weight
    # новий канал середнім по першим 3 або нулями
    new_weight[:, 3:4, :, :] = first_weight.mean(dim=1, keepdim=True)
ms[first_key] = new_weight
# load back
model.load_state_dict(ms)
torch.save({'model': model.state_dict()}, 'yolov5s_rgbd.pt')
```

Тренування моделі.

Для RGBD виконано partial fine-tune з однаковими гіперпараметрами:

- backbone: yolov5s (pretrained)
- розмір зображення: 640 × 640
- batch size: 16
- epochs: 30 і 50
- optimizer: SGD (momentum=0.937) / або AdamW
- lr: 0.01 (SGD) або 1e-3 (AdamW) з warmup

```
python train.py \
  --img 640 \
  --batch 16 \
  --epochs 50 \
  --data coco_rgbd.yaml \
  --weights yolov5s_rgbd.pt
```

У багатьох великих датасетах для комп'ютерного зору, таких як MS COCO, спостерігається класовий дисбаланс, коли деякі класи об'єктів представлені значно частіше за інші. Наприклад, клас «людина» в COCO зустрічається на багатьох зображеннях і складає близько 25-30% всіх об'єктних інстанцій, тоді як інші класи, такі як «тостер» чи «фен», мають дуже мало зображень. Це явище, яке називається class imbalance (класовий дисбаланс), призводить до того, що моделі комп'ютерного зору зазвичай добре вивчають класи з великою кількістю прикладів, але мають проблеми з точним розпізнаванням рідкісних або мало представлених класів. Враховуючи складність роботи з великими датасетами, було прийнято рішення використовувати тисяча зображень в якості початкового набору даних для навчання моделі. Такий розмір дозволяє зменшити навантаження на модель та зменшити час тренування без значних втрат у якості результату

для частих классів (зокрема клас людину, на якому буде проводитись перевірка моделі).’

4.2 Перевірка роботи моделі на власному датасеті

Для перевірки роботи моделі необхідно створити власний датасет, використовуючи модель MiDaS. Першим кроком буде запуск моделі локально, що включає в себе встановлення всіх необхідних бібліотек та середовищ. Після цього, фотографія (рис 4.2), в датасеті має бути пропущена через модель MiDaS для подальшої обробки, після якої фотографія буде виглядати як карта глибини (рис 4.3).

Рисунок 4.2 — Початкове зображення для створення карти глибини

Рисунок 4.3 — Зображення після обробки мережею MiDaS

Фото результати та їх виведення при виконанні завдання детекції на стандартній моделі RGB на власному датасеті (рис 4.4):

```
Class: person, Box: [319.89544677734375, 275.3390197753906,
341.9290466308594, 355.6991271972656], New Confidence: 0.61
Class: person, Box: [355.0561218261719, 277.4158935546875,
383.706298828125, 357.84918212890625], New Confidence: 0.47
Class: person, Box: [660.9616088867188, 287.2187805175781,
691.7142333984375, 348.08160400390625], New Confidence: 0.43
Class: person, Box: [425.02728271484375, 294.58612060546875,
441.11163330078125, 346.72509765625], New Confidence: 0.34
```

Рисунок 4.4 — Результат стандартної моделі

Фото результати та їх виведення при виконанні завдання детекції на мультимодальної моделі RGB-D на власному датасеті (рис 4.5):

Class: person, Box: [319.89544677734375, 275.3390197753906, 341.9290466308594, 355.6991271972656], Confidence: 0.73
Class: person, Box: [355.0561218261719, 277.4158935546875, 383.706298828125, 357.84918212890625], Confidence: 0.72
Class: person, Box: [660.9616088867188, 287.2187805175781, 691.7142333984375, 348.08160400390625], Confidence: 0.58
Class: person, Box: [425.02728271484375, 294.58612060546875, 441.11163330078125, 346.72509765625], Confidence: 0.44
Class: person, Box: [418.2667236328125, 294.2995910644531, 440.416748046875, 346.74163818359375], Confidence: 0.40
Class: person, Box: [416.9227600097656, 295.60089111328125, 430.54656982421875, 347.0309143066406], Confidence: 0.31

Рисунок 4.5 — Результат мультимодальної моделі

Код на виведення результатів:

```
# отримуємо bounding boxes
boxes = results[0].boxes.xyxy  # bounding box у вигляді [x1, y1, x2, y2]
confidences = results[0].boxes.conf  # ймовірності для кожного bounding box
class_ids = results[0].boxes.cls

# Вивіл
for box, conf, cls in zip(boxes, confidences, class_ids):
    print(f"Class: {results[0].names[int(cls)]}, Box: {box.tolist()}, Confidence: {conf:.2f}")
```

Для порівняння результаті також були використані метрики оцінювання ефективності моделей детекції об'єктів:

- **mAP50** (mean Average Precision at IoU=0.5) показує середню точність моделі при порозі перехрестя (IoU) 0.5. Це означає, що об'єкти вважаються правильно розпізнаними, якщо IoU з предсказаним об'єктом перевищує 50%.
- **mAP** (mean Average Precision) відображає середню точність моделі на різних порогах IoU (від 0.5 до 0.95) і дає загальну оцінку її роботи.
- **FPS** (Frames Per Second) вимірює швидкість роботи моделі, тобто кількість зображень, які вона може обробити за одну секунду.

  У таблиці наведені показники для двох варіантів YOLOv5: з кольоровими зображеннями (RGB) і з додатковою глибиною (RGBD).

Таблиця 4.1 — Метрики оцінювання

Model	Input	mAP50	mAP	FPS
YOLOv5 (RGB)	3 ch	0.37	0.23	110
YOLOv5 (RGBD)	4 ch	0.39	0.25	95

Значення mAP50 та mAP свідчать про те, що модель з RGBD зображеннями також має трохи кращу загальну точність.

FPS вказує на те, що модель, яка використовує зображення з додатковою глибиною (RGBD), працює трохи повільніше (менше FPS) порівняно з моделлю, яка використовує тільки кольорові зображення (RGB).

4.3 Висновок четвертого розділу

У практичному розділі було реалізовано повний цикл побудови мультимодальної системи розпізнавання об’єктів на основі глибинної інформації, згенерованої мережею MiDaS, та переднавченої моделі YOLOv5. Основною метою стало дослідження впливу додаткового каналу глибини (Depth) на якість детекції.

ВИСНОВКИ

У **першому розділі** здійснено глибокий аналіз предметної області мультимодального розпізнавання, досліджено сучасний стан розвитку методів, тенденції та ключові наукові підходи. Проведено огляд літератури, присвяченої методам інтеграції модальностей, особливостям роботи з RGBD-даними, сучасним архітектурам глибинних нейронних мереж та принципам побудови систем детекції. Також розглянуто існуючі інструменти, моделі та бібліотеки, що застосовуються у мультимодальних рішеннях. Результатом стало формування теоретичної бази та визначення ключових проблем, які потребують вирішення.

У **другому розділі** сформульовано технічні завдання, необхідні для побудови мультимодальної моделі: представлення даних, методи злиття модальностей, стратегії вирівнювання та узгодження інформації. Проаналізовано різні підходи до репрезентації сигналів, описано типи ф’южн-механізмів (early, mid, late fusion), а також алгоритми вирівнювання та подолання міжмодальних розбіжностей. Це дозволило побудувати концептуальну основу майбутньої архітектури та визначити, які технічні характеристики має враховувати практична реалізація.

У **третьому розділі** проведено розбір архітектурних рішень, датасетів та програмних інструментів. Як базову модель було обрано нейронну мережу YOLOv5, що забезпечує оптимальний баланс між точністю та швидкістю. Для формування карти глибини обґрунтовано використання моделі MiDaS, яка забезпечує високоякісну моно-стерео реконструкцію глибини. Також визначено набір бібліотек і фреймворків (PyTorch, OpenCV, Ultralytics), що забезпечують ефективну реалізацію та експериментальні дослідження.

У четвертому розділі виконано практичну реалізацію мультимодальної моделі, включаючи обробку даних, інтеграцію глибинного каналу та модифікацію архітектури YOLOv5 під early fusion.

Проведено навчання і тестування моделі на власному датасеті RGBD-зображень. Отримані експериментальні результати підтвердили, що використання додаткової модальності глибини підвищує точність детекції.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. A Survey of the Multi-Sensor Fusion Object Detection Task in Autonomous Driving [Електронний ресурс] // mdpi.com – Режим доступу: https://www.mdpi.com/1424-8220/25/9/2794
2. Approaches to Multimodal Fusion: Early, Intermediate, Late [Електронний ресурс] // apxml.com – Режим доступу: https://apxml.com/courses/intro-to-multimodal-ai/chapter-3-techniques-integrating-modalities/approaches-multimodal-fusion.
3. What are the challenges in building multimodal AI systems? [Електронний ресурс] // milvus.io – Режим доступу: https://milvus.io/ai-quick-reference/what-are-the-challenges-in-building-multimodal-ai-systems
4. A Survey of Modern Deep Learning based Object Detection Models [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2104.11892
5. End-to-End Object Detection with Transformers [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2005.12872
6. Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2309.04902
7. Quantum Inverse Contextual Vision Transformers (Q-ICVT): A New Frontier in 3D Object Detection for AVs [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2408.11207
8. What are some ethical concerns in multimodal AI systems? [Електронний ресурс] // milvus.io – Режим доступу: https://milvus.io/ai-quick-reference/what-are-some-ethical-concerns-in-multimodal-ai-systems
9. A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection [Електронний ресурс] // mdpi.com – Режим доступу: https://www.mdpi.com/1424-8220/25/1/214

10. 2D Object Detection: A Survey [Електронний ресурс] // mdpi.com – Режим доступу: https://www.mdpi.com/2227-7390/13/6/893
11. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2302.00389
12. What is DETR (Detection Transformers)? [Електронний ресурс] // roboflow.com – Режим доступу: https://blog.roboflow.com/what-is-detr/
13. Survey and Performance Analysis of Deep Learning Based Object Detection in Challenging Environments [Електронний ресурс] // roboflow.com – Режим доступу: https://doi.org/10.3390/s21155116
14. What is data augmentation? [Електронний ресурс] // ibm.com – Режим доступу: https://www.ibm.com/think/topics/data-augmentation
15. Deep Learning for Weakly-Supervised Object Detection and Object Localization: A Survey [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2105.12694
16. Multimodal object detection: an architecture using feature-level fusion and deep learning [Електронний ресурс] // springer.com – Режим доступу: https://link.springer.com/article/10.1007/s00521-025-11521-x
17. Learning Transferable Visual Models From Natural Language Supervision [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2103.00020
18. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2303.05499
19. Rethinking Multi-modal Object Detection from the Perspective of Mono-Modality Feature Learning [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/html/2503.11780v1
20. Towards Open-Vocabulary Multimodal 3D Object Detection with Attributes [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2508.16812

21. Building a Strong Pre-Training Baseline for Universal 3D Large-Scale Perception [Електронний ресурс] // emergentmind.com – Режим доступу: https://www.emergentmind.com/papers/2405.07201
22. Angelina Jolie? The humanoid robot Ameca, the person and machine connects - from trade fairs to museums he conquers the world [Електронний ресурс] // xpert.digital – Режим доступу: https://xpert.digital/en/the-humanoid-robot-ameca
23. Object Detection with Multimodal Large Vision-Language Models: An In-depth Review [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/html/2508.19294v1
24. Object Detection with Multimodal Large Vision-Language Models: An In-depth Review [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/html/2508.19294v1
25. What are some multimodal AI tools available for developers? [Електронний ресурс] // milvus.io – Режим доступу: https://milvus.io/ai-quick-reference/what-are-some-multimodal-ai-tools-available-for-developers
26. How Multimodal AI Enhances Quality Control Across Manufacturing Lines? [Електронний ресурс] // bluebash.co – Режим доступу: https://www.bluebash.co/blog/multimodal-ai-in-manufacturing-quality-control/
27. Developing reliable AI tools for healthcare [Електронний ресурс] // deepmind.google – Режим доступу: https://deepmind.google/discover/blog/codoc-developing-reliable-ai-tools-for-healthcare/
28. NVIDIA AGX Systems [Електронний ресурс] // nvidia.com – Режим доступу: https://www.nvidia.com/en-us/deep-learning-ai/products/agx-systems

29. NVIDIA Scores Consecutive Win for End-to-End Autonomous Driving Grand Challenge at CVPR [Електронний ресурс] // nvidia.com – Режим доступу: https://blogs.nvidia.com/blog/auto-research-cvpr-2025/
30. A Smart Security System Using Multimodal Features from Videos [Електронний ресурс] // springer.com – Режим доступу: https://link.springer.com/article/10.1134/S1054661819010218
31. Introducing Vulcan: Amazon's first robot with a sense of touch [Електронний ресурс] // aboutamazon.com – Режим доступу: https://www.aboutamazon.com/news/operations/amazon-vulcan-robot-pick-stow-touch
32. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1908.02265
33. Learning Transferable Visual Models From Natural Language Supervision [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2103.00020
34. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2205.11487
35. Learning Representations from Audio-Visual Spatial Alignment [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2011.01819
36. Vision–language representation learning with breadth and depth attention pre-training [Електронний ресурс] // sciencedirect.com – Режим доступу: https://www.sciencedirect.com/science/article/abs/pii/S0950705124015752
37. Multimodal Generative Models for Scalable Weakly-Supervised Learning [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1802.05335

38. Audio-Visual Spatial Alignment Requirements of Central and Peripheral Object Events [Електронний ресурс] // researchgate.net – Режим доступу: https://www.researchgate.net/publication/341316957_Audio-Visual_Spatial_Alignment_Requirements_of_Central_and_Peripheral_Object_Events
39. LXMERT: Learning Cross-Modality Encoder Representations from Transformers [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1908.07490
40. Proxy Convexity: A Unified Framework for the Analysis of Neural Networks Trained by Gradient Descent [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2106.13792
41. Multimodal Machine Learning: A Survey and Taxonomy [Електронний ресурс] // researchgate.net – Режим доступу: https://www.researchgate.net/publication/317185818_Multimodal_Machine_Learning_A_Survey_and_Taxonomy
42. mmFUSION: Multimodal Fusion for 3D Objects Detection [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2311.04058
43. FaceNet: A Unified Embedding for Face Recognition and Clustering [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1503.03832
44. Representation Learning with Contrastive Predictive Coding [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1807.03748
45. FaceNet - Using Facial Recognition System [Електронний ресурс] // geeksforgeeks.org – Режим доступу: https://www.geeksforgeeks.org/machine-learning/facenet-using-facial-recognition-system/
46. Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/2005.10242

47. Ultralytics YOLOv5 Architecture [Електронний ресурс] // ultralytics.com – Режим доступу: https://docs.ultralytics.com/yolov5/tutorials/architecture_description/
48. What is YOLOv5: A deep look into the internal features of the popular object detector [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/html/2407.20892v1
49. What is fine-tuning? [Електронний ресурс] // ibm.com – Режим доступу: https://www.ibm.com/think/topics/fine-tuning
50. How transferable are features in deep neural networks? [Електронний ресурс] // arxiv.org – Режим доступу: https://arxiv.org/abs/1411.1792
51. What Is Fine-Tuning? [Електронний ресурс] // coursera.org – Режим доступу: https://www.coursera.org/articles/what-is-fine-tuning
52. COCO Dataset [Електронний ресурс] // ultralytics.com – Режим доступу: https://docs.ultralytics.com/datasets/detect/coco/#sample-images-and-annotations
53. Cocodataset Overview [Електронний ресурс] // cocodataset.org – Режим доступу: https://cocodataset.org/#download
54. Model Hub MiDaS [Електронний ресурс] // pytorch.org – Режим доступу: https://pytorch.org/hub/intelisl_midas_v2/
55. A Guide to Data Labeling for Fine-tuning LLMs [Електронний ресурс] // gdsonline.tech – Режим доступу: https://www.gdsonline.tech/data-labeling-for-fine-tuning-llms/
56. Ultralytics Integrations [Електронний ресурс] // aboutamazon.com – Режим доступу: https://docs.ultralytics.com/ru/integrations/#deployment-integrations