Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент групи КН-22-1* — Підпис — *Ростислав ДИДО*
Група виконавця — Ім'я, ПРІЗВИЩЕ

Керівник: *Ph.D., ст. викл., каф. КН* — Підпис — *Олена СОБКО*
Науковий ступінь, посада — Ім'я, ПРІЗВИЩЕ

Нормоконтроль: *к.т.н., доц. каф. КН* — Підпис — *Руслан БАГРІЙ*
Науковий ступінь, посада — Ім'я, ПРІЗВИЩЕ

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор — Підпис — *Олександр БАРМАК*
Ім'я, ПРІЗВИЩЕ

17 червня 2026 р.

Хмельницький 2026

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

_____
(підпис)
д.т.н., професор Олександр БАРМАК
«*22*» *січня* *2026* року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання»*

2. Завдання видано студенту *Ростиславу Дидо*
(Ім'я, прізвище)

3. Керівник роботи *старший викладач кафедри КН Олена Собко*
(посада, ім'я, прізвище)

4. Затверджено наказом університету від «*20*» *січня* *2026* р. № *7*

5. Дата видачі завдання студенту: «*22*» *січня* *2026* р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Метою кваліфікаційної роботи бакалавра є підвищення якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, що полягає у підвищенні точності детекції та класифікації об'єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу. Для досягнення поставленої мети необхідно: дослідити проблеми орієнтації та безпеки осіб із порушеннями зору в міському середовищі; формалізувати задачу формування аудіопотоку доповненої реальності за відеоданими; розробити метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання; розробити інтелектуальну інформаційну систему для дослідження створеного методу; здійснити експериментальне дослідження методу шляхом оцінки точностідетекції та класифікації об'єктів, розпізнавання іменованих осіб, а також якості формування аудіопотоку.

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження теми кваліфікаційної роботи з керівником, складання календарного графіка виконання*	*січень 2026*	Виконано
2	*Ознайомлення з предметною областю, формулювання мети і задач дослідження, визначення об'єкта та предмета дослідження*	*лютий 2026*	Виконано
3	*Проєктування методу розв'язання задачі, опис архітектурних рішень, розроблення математичних моделей та алгоритмів.*	*березень 2026*	Виконано
4	*Обґрунтування інструментарію розробки, програмна реалізація розробленого методу, проведення експериментального тестування та оцінювання ефективності.*	*квітень 2026*	Виконано
5	*Написання тексту кваліфікаційної роботи, урахування зауважень керівника, оформлення згідно з вимогами*	*травень 2026*	Виконано
6	*Розроблення презентаційних матеріалів та попередній захист кваліфікаційної роботи*	*травень 2026*	Виконано
7	*Отримання відгуку керівника, рецензії, перевірка тексту кваліфікаційної роботина плагіат, нормоконтроль*	*червень 2026*	Виконано
8	*Підготовка до захисту та захист кваліфікаційної роботи*	*червень 2026*	Виконано

Виконавець: *студент групи КН-22-1* (Група виконавця) _____ (Підпис) *Ростислав ДИДО* (Ім'я, ПРІЗВИЩЕ)

Керівник: *Ph.D., ст. викл., каф. КН* (Науковий ступінь, посада) _____ (Підпис) *Олена СОБКО* (Ім'я, ПРІЗВИЩЕ)

Анотація

Тема кваліфікаційної роботи бакалавра: *«Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання»*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-22-1 Ростислав Дидо*

Керівник кваліфікаційної роботи бакалавра: *Ph.D., ст. викл., каф. КН Олена Собко*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерелін формації	
78	18	14	62	5

Метою кваліфікаційної роботи бакалавра є підвищення якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, що полягає у підвищенні точності детекції та класифікації об'єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу. Для реалізації запропонованого у роботі підходу використано методи комп'ютерного зору та глибокі згорткові нейромережі, що забезпечують розпізнавання об'єктів, ідентифікацію іменованих осіб і врахування пріоритетності потенційно небезпечних об'єктів у реальному часі.

Практичне значення роботи полягає у можливості використання розробленого методу формування аудіопотоку доповненої реальності на основі аналізу відеоданих із застосуванням глибоких згорткових нейронних мереж для створення інтелектуальних систем підтримки орієнтації та навігації осіб із порушеннями зору.

Ключові слова: openCV, CNN, доповнена реальність, аудіонавігація.

Виконавець: *студент групи КН-22-1* (Група виконавця) _____ (Підпис) *Ростислав ДИДО* (Ім'я, ПРІЗВИЩЕ)

Зміст

Перелік скорочень

Скорочення, термін, позначення	Пояснення
TTS	Text-to-Speech
GDPR	GeneralDataProtectionRegulation
BGR	Blue–Green–Red
CNN	ConvolutionalNeuralNetwork
COCO	CommonObjectsinContext
FPS	FramesPerSecond
GDPR	GeneralDataProtectionRegulation
IoU	IntersectionoverUnion
mAP	meanAveragePrecision
MOS	MeanOpinionScore
ONNX	OpenNeuralNetwork Exchange
VGG	VisualGeometryGroup
NMS	Non-MaximumSuppression
ЄС	Європейський Союз
ДР	доповнена реальність
ToF	Time-of-Flight
ReLU	RectifiedLinearUnit
FP	FalsePositive
FN	FalseNegative
DETR	DEtectionTRansformer

Вступ

Кваліфікаційна робота бакалавра присвячена підвищенню якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, що полягає у підвищенні точності детекції та класифікації об’єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу.

Актуальність. Актуальність дослідження зумовлена зростанням потреби у створенні інтелектуальних систем комп’ютерного зору та доповненої реальності, орієнтованих на підвищення рівня автономності та безпеки осіб із порушеннями зору в умовах складного міського середовища. Сучасні підходи до аналізу відеопотоку на основі глибоких згорткових нейронних мереж забезпечують високу ефективність у задачах детекції та класифікації об’єктів, однак не завжди забезпечують повноцінне семантичне представлення сцени у формі, придатній для аудіального сприйняття.

Водночас існує науково-практична проблема перетворення результатів відеоаналізу, зокрема ідентифікації та класифікації об’єктів і розпізнавання іменованих осіб, у структурований аудіопотік доповненої реальності, який забезпечує своєчасне та пріоритетне інформування користувача про об’єкти навколишнього середовища. Недостатня розробленість методів інтеграції модулів комп’ютерного зору з аудіальними інтерфейсами реального часу обумовлює необхідність створення відповідних моделей та алгоритмів, здатних забезпечити семантичну узгодженість, повноту та оперативність представлення інформації.

Таким чином, розроблення методу формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання є актуальним завданням, що поєднує задачі комп’ютерного зору, мультимодальної обробки даних та інтерфейсів людина-комп’ютер.

Об’єктом дослідження є процес формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору.

Предметом дослідження є методи глибокого навчання для детекції та класифікації об’єктів, розпізнавання іменованих осіб та формування аудіопотоку доповненої реальності на основі відеоданих.

Метою роботи є підвищення якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, що полягає у підвищенні точності детекції та класифікації об’єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу.

Для досягнення поставленої мети у роботі необхідно виконати наступні задачі: дослідити проблеми орієнтації та безпеки осіб із порушеннями зору в міському середовищі; формалізувати задачу формування аудіопотоку доповненої реальності за відеоданими; розробити метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання; розробити інтелектуальну інформаційну систему для дослідження створеного методу; здійснити експериментальне дослідження методу шляхом оцінки точності детекції та класифікації об’єктів, розпізнавання іменованих осіб, а також якості формування аудіопотоку.

Практичне значення роботи полягає у можливості використання розробленого методу формування аудіопотоку доповненої реальності на основі аналізу відеоданих із застосуванням глибоких згорткових нейронних мереж для створення інтелектуальних систем підтримки орієнтації та навігації осіб із порушеннями зору.

Отже, тема кваліфікаційної роботи бакалавра є актуальною як з наукової, так і з прикладної точки зору, оскільки спрямована на розв’язання соціально значущої задачі із застосуванням сучасних методів інтелектуального аналізу даних та розробленням програмного засобу, придатного до практичного використання.

Розділ 1 Аналіз предметної області формування аудіопотоку доповненої реальності на основі відеоданих для осіб із порушеннями зору

1.1 Аналіз інформаційних моделей доповненої реальності на основі відеоданих для осіб із порушеннями зору

Станом на сьогодні у світовому масштабі чисельність осіб із повною втратою зору становить близько 39 мільйонів, причому цей показник має тенденцію до щорічного зростання, що зумовлено процесами старіння населення, а також недостатньою доступністю медичних послуг у країнах, що розвиваються [1]. В Україні офіційно узагальнені статистичні дані щодо чисельності осіб із порушеннями зору відсутні, однак, за різними оцінками, кількість таких осіб перебуває в межах від 70 до 300 тисяч [2], Водночас, з огляду на умови воєнної агресії, кількість людей із порушеннями зору має тенденцію до подальшого зростання.

Попри впровадження на державному рівні низки ініціатив, спрямованих на розширення інклюзивності для осіб із порушеннями зору, зокрема облаштування міського простору тактильними елементами, підвищення доступності освітніх можливостей, реалізацію програм підтримки зайнятості та забезпечення правового захисту, наявних заходів усе ще недостатньо для повноцінного розв’язання цієї проблеми [3]. Сутність проблеми полягає у недостатній доступності інформаційних технологій, браку сучасних технічних рішень, що забезпечують інтеграцію осіб із порушеннями зору в повсякденну діяльність, а також у нерівномірному впровадженні інклюзивних підходів у різних регіонах [4].

У зв’язку з цим доцільним є доповнення державних ініціатив шляхом розроблення та впровадження інформаційних систем, здатних використовувати аудіопотоки доповненої реальності для суттєвого підвищення рівня автономності та якості життя осіб із порушеннями зору. Застосування подібних технологічних рішень сприятиме формуванню більш інклюзивного соціального середовища, у якому враховуються потреби людей з інвалідністю.

Розроблення та впровадження відповідного програмного забезпечення корелює із глобальною концепцією Цілей сталого розвитку Організації Об'єднаних Націй, які визначають стратегічні орієнтири соціально-економічного прогресу до 2030 року та спрямовані на забезпечення добробуту населення, подолання нерівності та формування стійкого середовища існування [5]. У цьому контексті запропоноване програмне рішення може розглядатися як інструмент цифрової трансформації, що забезпечує інтеграцію інноваційних технологій у процеси управління, доступу до послуг і соціальної взаємодії.

Ціль 3 «Міцне здоров'я і благополуччя» [6] передбачає забезпечення здорового способу життя та підвищення якості медичних послуг для всіх вікових груп. Реалізація програмного забезпечення сприяє досягненню цієї мети шляхом автоматизації процесів моніторингу стану здоров'я, забезпечення доступу до цифрових медичних сервісів, а також підвищення оперативності обробки даних. Це дозволяє оптимізувати прийняття управлінських рішень у сфері охорони здоров'я та забезпечити більш рівномірний доступ населення до якісних послуг.

Ціль 10 «Скорочення нерівності» [7] орієнтована на зменшення соціально-економічних диспропорцій як всередині країн, так і на глобальному рівні. Впровадження програмного забезпечення забезпечує рівний доступ до інформаційних ресурсів і цифрових сервісів незалежно від територіального розташування чи соціального статусу користувачів. Це створює передумови для цифрової інклюзії, розширення можливостей різних груп населення та зменшення бар'єрів у доступі до послуг.

Ціль 11 «Сталий розвиток міст і спільнот» [8] спрямована на формування безпечних, відкритих та екологічно стійких населених розділів. Розроблене програмне забезпечення може бути використане для оптимізації міської інфраструктури, підвищення ефективності управління ресурсами та забезпечення інтеграції цифрових сервісів у міське середовище. Застосування таких рішень сприяє підвищенню якості життя населення, розвитку «розумних» міст та підвищенню екологічної відповідальності.

Отже, використання сучасних інформаційних технологій, зокрема методів штучного інтелекту, є доцільним інструментом вирішення окресленої проблеми,

оскільки дозволяє забезпечити комплексний підхід до підвищення ефективності управління, покращення доступу до послуг та реалізації принципів сталого розвитку в умовах цифрової трансформації суспільства.

1.2 Огляд методів комп'ютерного зору та підходів до формування аудіопотоку доповненої реальності на основі відеоданих

У розробленні програмного засобу для підтримки незрячих користувачів важливо оцінювати технології не лише за точністю, а й за швидкодією, вимогами до ресурсів, можливістю роботи в реальному часі та простотою інтеграції. Для детекції об'єктів одним із найпрактичніших рішень є моделі сімейства YOLO, які використовують одноетапний підхід, поєднуючи локалізацію та класифікацію в одному проході нейронної мережі. Це забезпечує низьку затримку обробки кадрів і придатність до застосування в асистивних системах реального часу. Завдяки балансу між точністю, швидкодією YOLO доцільно використовувати як основу для розпізнавання об'єктів [9].

Альтернативою є трансформерні моделі реального часу, зокрема RT-DETR. Вони ефективніше враховують глобальні зв'язки між елементами зображення та можуть покращувати якість детекції у складних сценах. Однак такі рішення зазвичай потребують більше обчислювальних ресурсів, тому для доступних систем їх доцільніше розглядати як перспективний напрям розвитку, а не базове рішення [10]. Для мобільних застосунків можуть використовуватися спеціалізовані SDK, наприклад Google ML KitObjectDetectionandTracking, який оптимізований для локальної обробки на пристрої. Проте він має меншу гнучкість налаштування порівняно з YOLO та подібними моделями [11].

Для зменшення надлишкових повідомлень доцільно використовувати алгоритми трекінгу об'єктів, зокрема ByteTrack або BoT-SORT. Вони дозволяють відстежувати об'єкти між кадрами та повідомляти користувача лише про появу нових об'єктів, або потенційну небезпеку [12].

Для формування аудіоповідомлень можуть застосовуватися системи синтезу мовлення. Для локальної реалізації доцільною є бібліотека pyttsx3, яка

не потребує підключення до інтернету та забезпечує автономність роботи, хоча поступається сучасним хмарним рішенням за природністю звучання [13]. Альтернативою є Edge TTS, який використовує голосові моделі Microsoft і забезпечує вищу якість озвучення, але залежить від мережевого з'єднання [14].

Перспективними напрямами розвитку є використання просторового звуку та соніфікації. Вони дають змогу передавати не лише факт наявності об'єкта, а й інформацію про його напрямок, відстань або наближення за допомогою звукових сигналів. Це може зменшити кількість голосових повідомлень і підвищити зручність взаємодії, однак потребує додаткових досліджень і тестування з користувачами [15, 16].

З огляду на проведений аналіз, для реалізації програмного засобу доцільно використати модель сімейства YOLO як основний засіб розпізнавання об'єктів у відеопотоці, оскільки вона забезпечує необхідний баланс між швидкодією, точністю та складністю інтеграції. Для формування аудіоповідомлень доцільним є застосування локального синтезу мовлення, що забезпечує автономність роботи системи та не потребує постійного доступу до мережі. Таким чином, найбільш обґрунтованим для базової реалізації є поєднання YOLO, модуля фільтрації результатів детекції та локального TTS-компонента, оскільки така архітектура відповідає вимогам реального часу, доступності, автономності та практичної корисності для незрячого користувача.

1.3 Аналіз підходів до розпізнавання об'єктів у відеопотоці та формування аудіопотоку доповненої реальності

Підвищення рівня безпеки осіб із порушеннями зору із застосуванням сучасних інформаційних технологій, засобів та методів штучного інтелекту є актуальним напрямом наукових досліджень.У таких системах методи штучного інтелекту використовуються для розпізнавання об'єктів, аналізу контексту сцени та визначення релевантної для користувача інформації. Частина науковців зосереджує увагу на інтеграції багатосенсорних даних, у межах якої поєднуються візуальна інформація, аудіосигнали та дані з різноманітних

сенсорних пристроїв з метою формування більш повного та структурованого уявлення про навколишнє середовище. Подібні системи, як правило, забезпечують адаптивний аудіозворотний зв'язок, який динамічно враховує індивідуальні потреби користувача, а також змінні умови зовнішнього середовища.

Автори [17] розробли голосовий мобільний застосунок для смартфонів, призначений для підтримки осіб із порушеннями зору під час ідентифікації об'єктів та орієнтації у навколишньому середовищі. Запропонована система поєднує функції розпізнавання текстових фрагментів, виявлення людей і різноманітних об'єктів, а також використовує сенсорні модулі для детекції перешкод. Для аналізу візуальної інформації застосовується алгоритм K-nearestneighbor, який забезпечує класифікацію зображень та сприяє формуванню відповідних голосових повідомлень для користувача.

У науковій статті представлено автоматизовану систему ідентифікації об'єктів, призначену для підтримки осіб із порушеннями зору, у якій для виявлення та локалізації об'єктів застосовано моделі глибокого навчання RFCN та Mask R-CNN [18]. Найвищу ефективність продемонструвала модель RFCN, для якої значення середньої точної відповідності (mAP) становило 0,825. У межах розробленої системи використано обчислювальну платформу RaspberryPi, до якої під'єднано камеру та ультразвуковий сенсор, що забезпечує вимірювання відстані до об'єктів. Отримана інформація обробляється системою та передається користувачеві у формі звукового зворотного зв'язку.

Автори [19] запропонували інноваційний підхід до забезпечення доступу осіб із порушеннями зору до візуальної інформації шляхом перетворення зображень на звукові описи. Розроблена система ґрунтується на використанні гібридної моделі, що поєднує архітектуру VGG-16, згорткові нейронні мережі (CNN) та методи обробки природної мови (NLP). Така комбінація алгоритмів забезпечила точність розпізнавання об'єктів і формування аудіоописів на рівні 0,9580 . Для навчання моделі було використано набір даних, що містить понад 8 000 зображень із відповідними назвами об'єктів та аудіопоясненнями.

Запропоновано авторами статті [20] представили систему перетворення візуальної інформації у мовні повідомлення для осіб із порушеннями зору, яка поєднує використання бібліотеки OpenCV та алгоритму YOLO для ідентифікації об’єктів і подальшого формування аудіоописів. Запропоноване рішення продемонструвало високий рівень точності, що становить 0,96. Водночас у межах проведеного дослідження розглянуто розпізнавання відносно обмеженої кількості об’єктів навколишнього середовища, що дещо звужує можливості практичного застосування системи.

У статті [21] описано систему розпізнавання об’єктів у режимі реального часу, що базується на використанні алгоритму YOLO, оптимізованого для функціонування на мобільних пристроях. Запропоноване рішення надає можливість особам із порушеннями зору ідентифікувати об’єкти у навколишньому просторі за допомогою текстових і звукових сповіщень, що сприяє підвищенню їхньої автономності та зменшує залежність від сторонньої допомоги або спеціалізованих технічних засобів.

Аналіз сучасних наукових досліджень дає підстави стверджувати, що методи та засоби штучного інтелекту широко застосовуються для підвищення рівня безпеки осіб із порушеннями зору. Водночас запропоновані підходи мають певні обмеження, серед яких варто відзначити обмежену кількість об’єктів, що підлягають розпізнаванню, а також недостатньо високі показники точності їхньої ідентифікації. Такі недоліки знижують ефективність використання подібних систем у контексті забезпечення безпеки користувачів. У зв’язку з цим виникає потреба у розробленні методу формування аудіопотоку доповненої реальності за відеоданими для людей із порушеннями зору із застосуванням засобів глибокого навчання. Такий метод має бути спрямований на підвищення рівня інклюзії користувачів шляхом автоматизованого розпізнавання об’єктів у навколишньому середовищі, інтерпретації результатів відеоаналізу та їх подальшого перетворення у доступні аудіальні повідомлення. Це дає змогу забезпечити користувача релевантною інформацією про об’єкти та потенційні перешкоди, що може сприяти підвищенню автономності, безпеки та ефективності повсякденної взаємодії з середовищем.

1.4 Етичні та правові аспекти розроблення інтелектуальних систем для осіб із порушеннями зору

Етичний і правовий аналіз програмних засобів для людей із порушеннями зору має враховувати не лише технологічні аспекти, а й питання прав людини, безпеки, доступності інформації та меж делегування рішень алгоритмам. Це дає змогу оцінювати систему з позицій недискримінації, автономності користувача, захисту персональних даних і відповідального використання ШІ [22].

Міжнародні та європейські підходи розглядають асистивні цифрові рішення як засіб розширення автономності користувача, а не заміну людського сприйняття та контролю. Тому програмний засіб, що генерує аудіоописи на основі відеоданих, повинен виконувати допоміжну функцію та підтримувати орієнтування користувача без підміни його остаточних рішень. Такий людиноцентричний підхід відповідає рекомендаціям UNESCO та європейським принципам регулювання AI, спрямованим на захист безпеки й прав [23].

Одним із ключових етичних питань є відповідальність за хибний аудіовивід, помилки розпізнавання об’єктів або ідентифікації осіб. Відповідно до міжнародних документів, система штучного інтелекту не є суб’єктом права, тому відповідальність завжди покладається на фізичних або юридичних осіб, залучених до її життєвого циклу [24]. UNESCO наголошує, що підзвітність не може бути передана машині [25]. У європейському підході розробник відповідає за архітектуру, якість даних, точність, безпеку та механізми людського контролю, а користувач або організація-оператор – за належне впровадження й експлуатацію системи [26].

Помилковий аудіовивід не повинен автоматично вважатися ризиком, який повністю несе користувач. Якщо шкода виникає через дефект програмного забезпечення, неналежне оновлення, помилки розпізнавання або відсутність попереджень про обмеження системи, відповідальність може наставати для відповідного економічного оператора. Європейська Директива про відповідальність за дефектну продукцію поширює цей режим і на програмне забезпечення [27]. Водночас під час правової оцінки враховуються шкода,

дефектність, причинний зв’язок та можливі недбалі дії самого користувача [28]. Тому відповідальність аналізується насамперед на етапах розроблення, випуску, супроводу та експлуатації системи, а не покладається на алгоритм [29].

З етичної точки зору програмний засіб має позиціонуватися як допоміжний інструмент, а не безпомилкова система навігації. Аудіоповідомлення повинні розглядатися як додатковий канал орієнтування, що не замінює обережність користувача та інші засоби мобільності. Тому архітектура системи повинна передбачати людський контроль, можливість втручання, інформування про межі точності та доступні інструкції з безпеки. Відповідно до вимог ЄС, виробники зобов’язані надавати зрозумілу та доступну інформацію про безпечне використання продукції, зокрема для осіб з інвалідністю [30]. Для асистивних систем це особливо важливо, оскільки недостатнє інформування про обмеження моделі саме по собі може створювати ризики [31].

Окремий правовий аспект пов’язаний з ідентифікацією людей. У такому разі обробляються біометричні дані, які GDPR відносить до спеціальних категорій персональних даних. Їх використання потребує законної підстави, мінімізації даних, контролю доступу, безпечного зберігання, журналювання та реалізації принципу *privacybydesign*. Функція ідентифікації осіб впливає на право на приватність, захист персональних даних і можливість оскарження помилкових рішень системи [32].

Отже, хибний аудіовивід або неточна класифікація мають розглядатися як результат взаємодії кількох рівнів відповідальності. Міжнародні та європейські норми визначають, що етична й правова відповідальність залишається за особами та організаціями, які розробляють, постачають, інтегрують, супроводжують і використовують систему. Для користувача ж така програма повинна залишатися допоміжним засобом орієнтування, що підвищує інклюзію та безпеку, але не є автономним джерелом безпомилкових рішень.

1.5 Мета, задачі та вимоги до реалізації інтелектуальної системи

Метою кваліфікаційної роботи бакалавра є підвищення інформативності та доступності сприйняття навколишнього середовища для осіб із порушеннями зору, що полягає у підвищенні точності виявлення та класифікації об’єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу.

Для досягнення мети потрібно виконати наступні задачі:

– дослідити проблеми орієнтації та безпеки осіб із порушеннями зору в міському середовищі;

– формалізувати задачу формування аудіопотоку доповненої реальності за відеоданими;

– розробити метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання;

– розробити інтелектуальну інформаційну систему для дослідження створеного методу;

– здійснити експериментальне дослідження методу шляхом оцінки точності детекції та класифікації об’єктів, розпізнавання іменованих осіб, а також якості формування аудіопотоку.

Розділ 2 Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання

2.1 Формалізація задачі формування аудіопотоку доповненої реальності

У межах кваліфікаційної роботи розв’язується задача формування аудіопотоку доповненої реальності за відеоданими, вирішення якої дозволить підвищити інформативність та доступність сприйняття навколишнього середовища особами із порушеннями зору. Задача полягає у побудові інформаційного перетворення F, що з відеопотоку V формує семантично впорядкований аудіопотікA із застосуванням проміжного семантичного представлення сцени S, утвореного на основі множини виявлених у кадрі об’єктів O. Композицію перетворень формально подамо у вигляді:

$$V \rightarrow O \rightarrow S \rightarrow A, \qquad (2.1)$$

де V – вхідний відеопотік, O – множина виявлених у кадрі об’єктів, S – семантичне представлення сцени, A – вихідний аудіопотік. Схема зображена на рисунку 2.1.

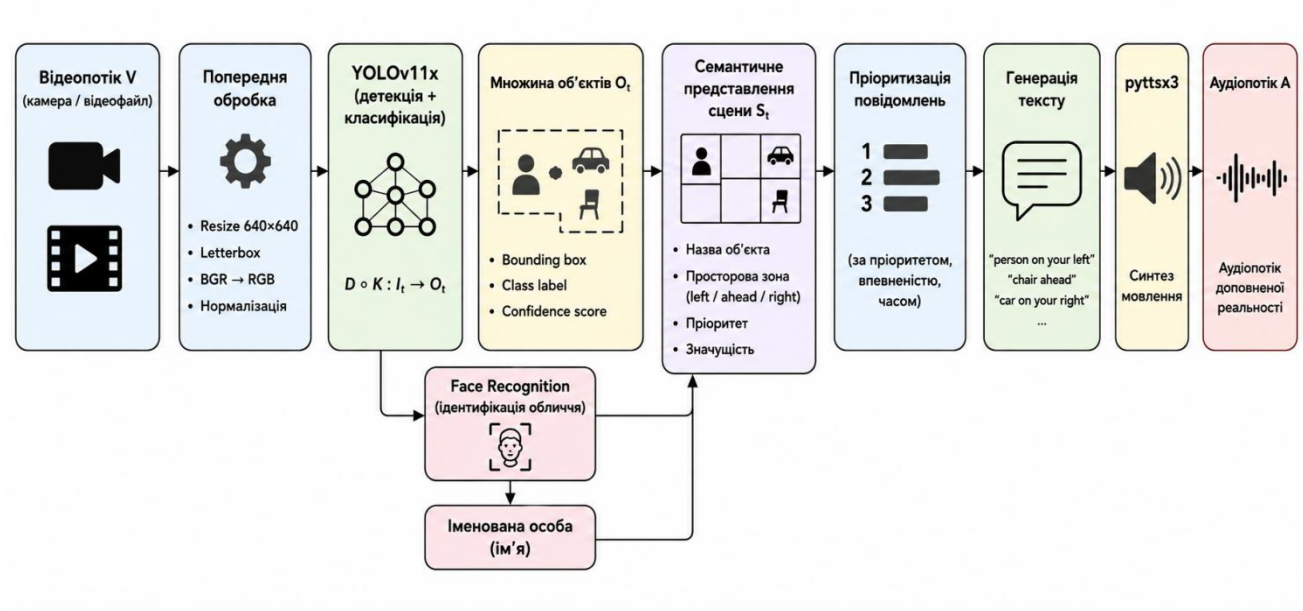


Рисунок 2.1 – Пайплайн формування аудіопотоку доповненої реальності

У роботі задача декомпонується на чотири підзадачі: формалізацію вхідних та вихідних даних (розділ 2.1.1), детекцію та класифікацію об’єктів

(розділ 2.1.2), розпізнавання та ідентифікацію іменованих осіб (розділ 2.1.3), побудову семантичного представлення сцени та визначення порядку озвучення (розділ 2.1.4).

2.1.1 Формалізація вхідних та вихідних даних

Вхідними даними методу є відеопотік, поданий упорядкованою у часі послідовністю кадрів:

$$V = \{ I_t \}, \ t = 1, \ldots, T, \tag{2.2}$$

де I_t– кадр відеопотоку в момент часу t, T – кількість опрацьованих кадрів. Кожний кадр I_t розглядається як тривимірний тензор

$$I_t \in R^{\{H \times W \times C\}}, \tag{2.3}$$

де H, W – висота та ширина кадру (у пікселях), C= 3 – кількість колірних каналів (RGB). При частоті надходження кадрів f кадр/с інтервал часу між сусідніми кадрами становить $\Delta t = 1/f$, що визначає часову роздільну здатність аналізу.

$$A = \{ a_j \}, \ j = 1, \ldots, M, \tag{2.4}$$

де M – кількість сформованих аудіальних повідомлень за інтервал спостереження. Кожне повідомлення описується трійкою

$$a_j = (\tau_j, text_j, \pi_j), \tag{2.5}$$

де τ_j – момент генерації повідомлення, $text_j$ – його текстовий вміст, π_j – пріоритет озвучення. Текст формується у форматі «семантична мітка + словесна ознака просторового положення», що забезпечує компактність повідомлення та узгоджується з можливостями офлайн-синтезатора мовлення pyttsx3.

2.1.2 Формалізація задачі детекції та класифікації об’єктів у відеопотоці

У межах методу детекція та класифікація утворюють єдину процедуру аналізу кадру (алгоритм 2.1), результатом якої є множина об’єктів із координатами, класовими мітками та оцінками впевненості. Формально задача детекції подається у вигляді відображення:

$$D : I_t \rightarrow B_t,\ B_t = \{ b_t^k \},\ k = 1, \ldots, n_t, \tag{2.6}$$

де n_t – кількість виявлених у кадрі об’єктів, b_t^k – обмежувальний прямокутник k-го об’єкта, поданий координатами [33]

$$b_t^k = (x_1^k, y_1^k, x_2^k, y_2^k) \tag{2.7}$$

у системі координат оригінального кадру з початком у верхньому лівому куті.

Класифікація ставить у відповідність кожному прямокутнику класову мітку та оцінку впевненості:

$$K : b_t^k \rightarrow (c_t^k, p_t^k),\ \ p_t^k \in [0, 1] \tag{2.8}$$

Композиція D ∘ K формує множину виявлених у кадрі об’єктів

$$O_t = \{ o_t^k \},\ \ o_t^k = (b_t^k, c_t^k, p_t^k, m_t^k), \tag{2.9}$$

деm_t^k– уточнена семантична мітка об’єкта; для більшості об’єктів $m_t^k = c_t^k$, а для об’єктів класу «person» уточнюється процедурою ідентифікації іменованих осіб (підрозділ 2.1.3). Узгодженість передбачення з еталонною розміткою оцінюється через показник *IoU*; його формальне означення подано у підрозділі 2.4.1.

Алгоритм 2.1 – Псевдокод детекції та класифікації об’єктів у кадрі

Вхідні дані: кадр I_t, навчена нейромережева модель $D * K$ (YOLOv11x), пороги впевненості τ_{conf} і $IoU\tau_{iou}$.
Вихідні дані: множина $O_t = \{ o_t^k \}$ виявлених об’єктів кадру (вираз 2.9).

```
Î_t ← Letterbox(Resize(I_t, 640 × 640)) / 255 // попередня обробка кадру
Y_t ← D(Î_t) // прямий прохід магістраллю та шийкою
(B_t, C_t, P_t) ← Decode(Y_t) // декодування виходів голови детектора (2.6)–(2.8)
(B_t, C_t, P_t) ← NMS(B_t, C_t, P_t, τ_iou) // придушення немаксимумів
(B_t, C_t, P_t) ← Filter(B_t, C_t, P_t, τ_conf, ALLOWED_CLASSES) // фільтрація за впевненістю та класовим простором
O_t ← ∅ // ініціалізація вихідної множини
Для кожного k = 1, …, n_t виконати: O_t ← O_t ∪ { ( b_t^k, c_t^k, p_t^k, c_t^k ) } // формування об’єктів за виразом (2.9)
Повернути O_t
```

Отже, у межах описаного методу детекція та класифікація розглядаються як єдиний етап аналізу кадру, що перетворює вхідне зображення на структуровану множину об’єктів із координатами, класовими мітками, оцінками впевненості та семантичними уточненнями, придатними для подальшої ідентифікації осіб і формування аудіоповідомлень.

2.1.3 Формалізація задачі розпізнавання та ідентифікації іменованих осіб

Окремою підзадачею у складі методу є розпізнавання та ідентифікація іменованих осіб серед об’єктів класу «person». Тут визначається не категорія об’єкта, а його ідентичність у межах наперед сформованої бази зареєстрованих користувачів. Задача розв’язується в ознаковому просторі ембедингів обличчя.

Для кожного прямокутника b_t^k класу person з кадру I_t вирізається область f_t^k, для якої будується ознаковий вектор:

$$e_t^k = E\,(\,f_t^k\,), \quad e_t^k \in R^{128}, \tag{2.10}$$

де E – функція побудови ембедингу обличчя (реалізована у пакеті face_recognition, модель типу ResNet-29, навчена з тріплетною функцією втрат). Міра близькості між поточним вектором e_t^k та i-м еталонним вектором $e^{\{(i)\}}$ з бази E_{ref}обчислюється за евклідовою нормою:

$$d\,(\,e_t^k, e^{(i)}\,) = ||\,e_t^k - e^{(i)}\,||_2. \tag{2.11}$$

Рішення про присвоєння виявленому обличчю імені відомої особи приймається за пороговим правилом:

$$name\,(\,o_t^k\,) = name^{(k^*)}, \;\; k^* = argmin_i\, d\,(\,e_t^k, e^{(i)}\,), \;\; d\,(\,e_t^k, e^{(k^*)}\,) < \tau, \tag{2.12}$$

де τ – поріг прийняття рішення про збіг особи (для пакета *face_recognition* типовим є значення τ = 0,48). У протилежному випадку об’єкту присвоюється загальна мітка m_t^k = *person* без ідентифікації імені.

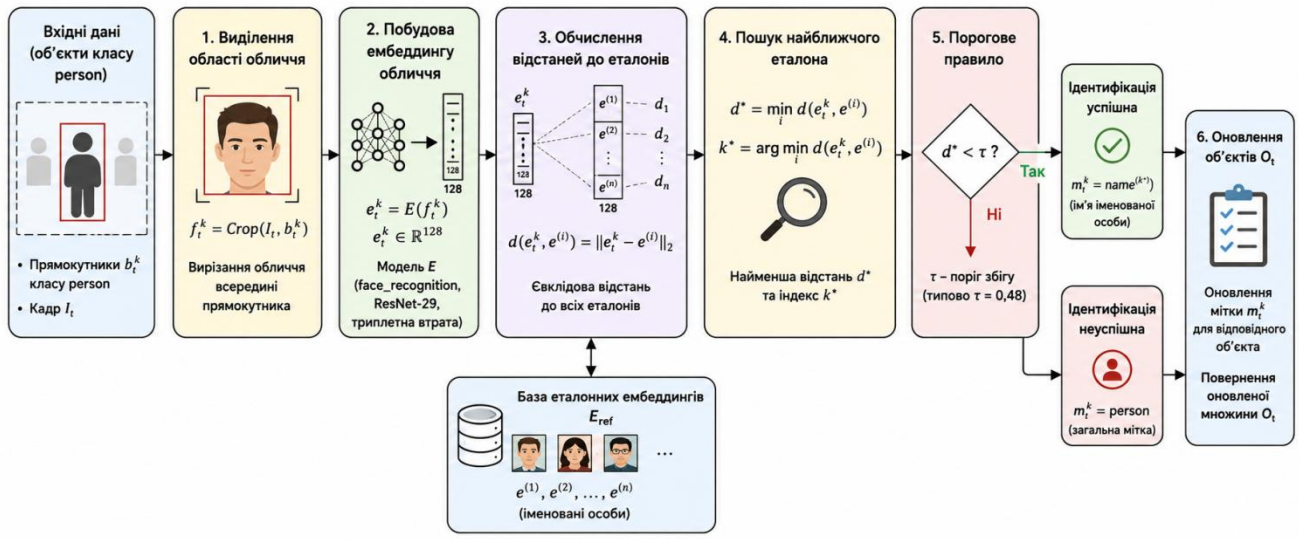


Рисунок 2.2 – Послідовність кроків ідентифікації обличчя та класифікації іменованої осіб у складі методу

Для зниження обчислювального навантаження ембединг будується раз на $EVERY_N$ кадрів із кешуванням відповідності між прямокутником та присвоєним іменем; у проміжних кадрах мітка успадковується з найближчого попереднього ідентифікованого вікна.

Алгоритм 2.2 – Псевдокод ідентифікації іменованих осіб у кадрі

Вхідні дані: множина обмежувальних прямокутників $B_t^{\{person\}} \subset B_t$, кадр I_t, база ембедингів $E_{ref} = \{e^{(i)}\}$, поріг τ, лічильник кадрів t.
Вихідні дані: оновлена множина об'єктів O_t із уточненими мітками m_t^k.

```
1. Якщо (t mod EVERY_N) ≠ 0: повернути O_t з кешу // пропуск кадру для зниження навантаження
2. Для кожного b_t^k ∈ B_t^{person} виконати кроки 3–8
3. f_t^k ← Crop(I_t, b_t^k) // виділення області обличчя
4. e_t^k ← E(f_t^k) // побудова 128-вимірного ембедингу (2.10)
5. d* ← min_i d(e_t^k, e^(i)), k* ← argmin_i d(e_t^k, e^(i)) // пошук найближчого еталона (2.11)
6. Якщо d* < τ: m_t^k ← name^(k*); інакше m_t^k ← «person» // порогове правило (2.12)
7. Cache[k] ← m_t^k // оновлення кешу
8. Оновити відповідний o_t^k у O_t
9. Повернути O_t
```

Отже, ідентифікація іменованих осіб забезпечує семантичне уточнення об'єктів класу «person» шляхом зіставлення ембедингів облич із базою еталонних векторів, що дає змогу замінити загальну мітку об'єкта на ім'я конкретної особи за умови виконання порогового критерію близькості.

2.1.4 Побудова семантичного представлення сцени, пріоритизація та порядок озвучення об'єктів з відеопотоку

Після виконання процедур детекції, класифікації та ідентифікації результати відеоаналізу об'єднуються у єдине семантичне представлення сцени через оператор Φ:

$$S_t = \Phi\,(O_t). \tag{2.13}$$

Кожному об'єкту o_t^k у S_t відповідає четвірка

$$s_t^k = (l_t^k, z_t^k, r_t^k, u_t^k), \tag{2.14}$$

де l_t^k – текстова мітка (отримана з c_t^k або з імені іменованої особи); z_t^k – словесна ознака просторового положення відносно користувача (left, ahead, right) за поділом ширини кадру на три рівні зони за центром обмежувального

прямокутника; r_t^k – пріоритет озвучення, заданий фіксованою таблицею пріоритетів для множини *ALLOWED_CLASSES*; u_t^k – ознака значущості об'єкта.

Ознака значущості u_t^k обчислюється як композиція критеріїв: належність c_t^k до пріоритетного класового простору, перевищення оцінкою p_t^k порогу впевненості τ_{conf} і перевищення відносною площею обмежувального прямокутника порогу присутності.

Порядок озвучення об'єктів визначається лексикографічним правилом: пріоритет r_t^k (зростання), оцінка p_t^k (спадання), момент появи об'єкта τ_t^k(зростання).

Для кожного значущого $s_t^k \in S_t$ текст повідомлення $text_j$ будується за шаблоном *«{l} onyour {z}»*. Сформоване представлення S_t є безпосереднім джерелом для модуля синтезу аудіальних повідомлень (розділ 2.2).

Алгоритм 2.3 – Псевдокод формування семантичного представлення сцени та черги озвучення

Вхідні дані: множина $O_t = \{ o_t^k \}$ (вираз 2.9), таблиця пріоритетів P, поріг впевненості τ_{conf}, поріг відносної площі τ_{area}, ширина кадру W.
Вихідні дані: впорядкована послідовність повідомлень $\{ a_j \}$ для синтезу мовлення.

1. $S_t \leftarrow \emptyset$ // ініціалізація
2. Для кожного $o_t^k \in O_t$ виконати кроки 3–7
3. $l_t^k \leftarrow (m_t^k \neq «person») \ ? \ m_t^k : c_t^k$ // вибір мітки з урахуванням ідентифікації
4. $z_t^k \leftarrow Zone(b_t^k, W)$ // left / ahead / right за центром b_t^k
5. $r_t^k \leftarrow P[l_t^k]$ // табличний пріоритет
6. $u_t^k \leftarrow [\ p_t^k \geq \tau_{conf}\ ]^{[\ Area(btk)/(H \bullet W) \geq \tau area\ ]}$// ознака значущості
7. Якщо $u_t^k = 1$: $S_t \leftarrow S_t \cup (l_t^k, z_t^k, r_t^k, u_t^k)$
8. Сортувати S_t за ключем $(r_t^k \uparrow, p_t^k \downarrow, \tau_t^k \uparrow)$ // лексикографічна пріоритизація
9. Для кожного $s_t^k \in S_t$ сформувати $text_j \leftarrow f«l_t^k$ on your $z_t^k»$ та $a_j \leftarrow (\tau_t, text_j, r_t^k)$ // композиція повідомлення (2.5)
10. Повернути впорядковану a_j

Отже, семантичне представлення сцени виконує роль проміжного рівня між результатами комп'ютерного зору та модулем озвучення, оскільки перетворює множину виявлених об'єктів на впорядковану послідовність текстових повідомлень з урахуванням їхньої мітки, просторового положення, впевненості та пріоритету для користувача.

2.2 Етапи методу формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання

Запропонований метод реалізує композицію $V \to O \to S \to A$ (вираз 2.1) у чотири послідовні етапи. На етапі попередньої обробки кадр I_t отримується з джерела (камера або файл), перетворюється з простору BGR у RGB і нормалізується відповідно до вимог нейромережевих моделей. На етапі детекції та класифікації виконується відображення $D \circ K$ (вирази 2.6–2.8) нейромережевою моделлю YOLOv11x, унаслідок чого формується множина об'єктів O_t. Для об'єктів класу person виконується додатковий етап ідентифікації іменованих осіб (вирази 2.10–2.12). На етапі побудови семантичного представлення сцени множина O_t перетворюється у впорядкований опис S_t за оператором Φ (вирази 2.13–2.14). На заключному етапі формується аудіопотікA: послідовність текстових повідомлень $text_j$ подається на офлайн-синтезатор мовлення pyttsx3; потоки мовлення та сигналізації керуються чергою з пріоритетом, побудованою за r_t^k.

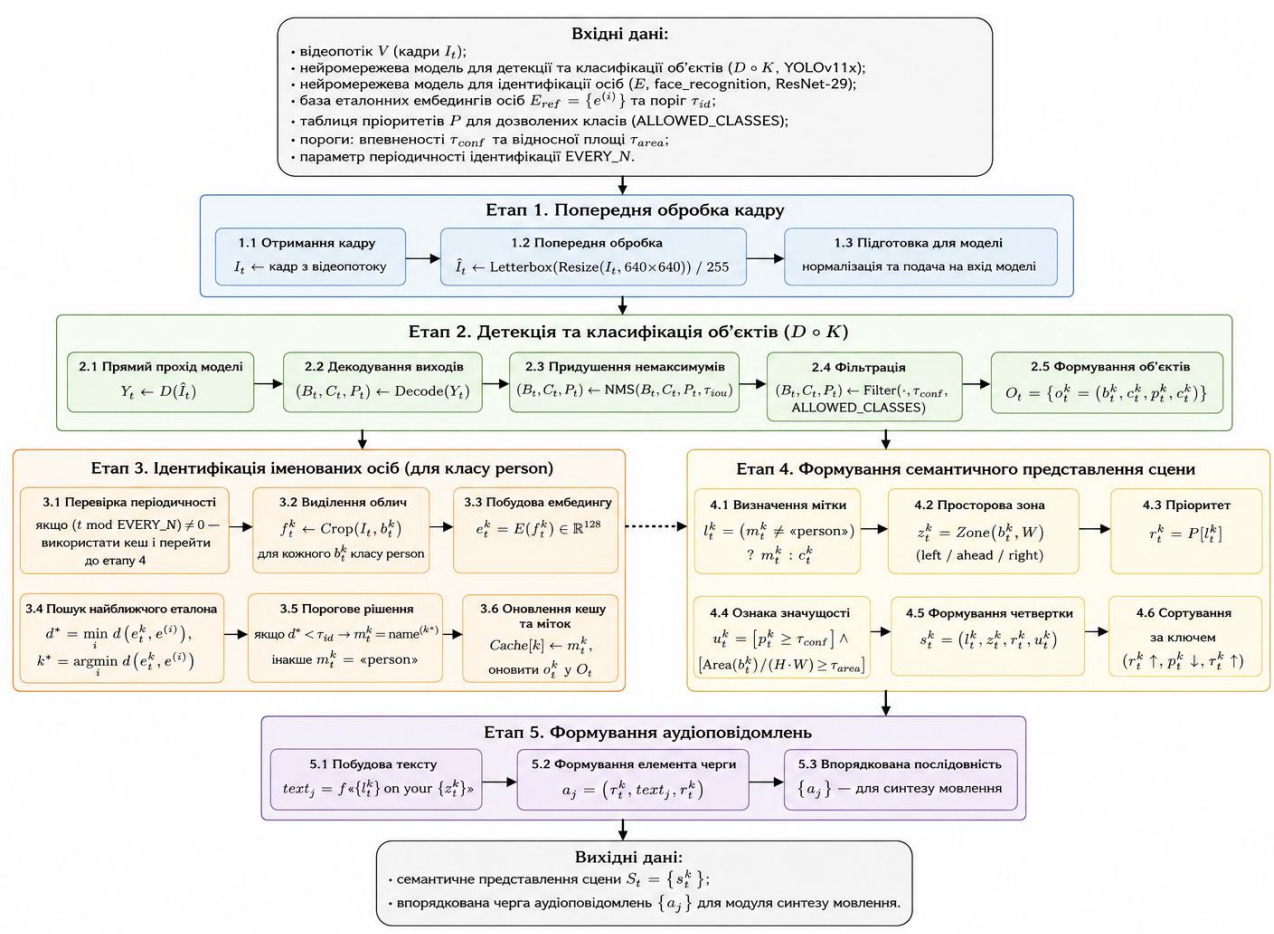


Рисунок 2.3 – Запропонований у роботі метод

Рисунок 2.3 узагальнює метод, формалізований у підрозділі 2.1, і відображає послідовність перетворень відеоданих відповідно до композиції $V \to O \to S \to A$. На схемі показано, як вхідний відеопотік через попередню обробку кадру I_t, детекцію та класифікацію $D \circ K_D$, ідентифікацію іменованих осіб, формування семантичного представлення $S_t = \Phi(O_t)$ і пріоритизацію повідомлень перетворюється на впорядкований аудіопотіка. Отже, схема не описує архітектуру програмного забезпечення, а подає узагальнену логіку методу з деталізацією його основних підкроків та відповідних математичних позначень.

2.3 Розроблення та навчання моделей глибокого навчання для аналізу відеопотоку

Для реалізації запропонованого методу необхідно отримати дві нейромережеві моделі. Перша – одноетапний детектор YOLOv11x – виконує детекцію та класифікацію об’єктів у кадрі (відображення $D \circ K$, вирази 2.6–2.8) і використовується з попередньо натренованими на COCO вагами. Друга – власна згорткова нейронна мережа (CNN) – виконує класифікацію іменованих осіб серед об’єктів класу person і донавчається на індивідуальній базі зображень користувачів. Допоміжно для побудови 128-вимірного ембедингу обличчя застосовується модель ResNet-29 з пакета *face_recognition* (вираз 2.10). Архітектуру та налаштування цих компонентів подано у підрозділах 2.3.1 і 2.3.2, формування наборів даних – у 2.3.3, процедуру навчання – у 2.3.4; зведений перелік моделей наведено у таблиці 2.1.

Таблиця 2.1 – Перелік нейромережевих моделей інтелектуальної системи

Призначення	Модель	Розмір входу	Вихід	Серіалізація
Детекція та класифікація об’єктів	YOLOv11x	640 × 640 × 3	80 класів COCO, координати B_t	.pt (Ultralytics)
Класифікація іменованих осіб	CNN-класифікатор	32 × 32 × 3	$L + 1$ клас (Δ^L)	.keras
Ембединг обличчя	face_recognition (ResNet-29)	150 × 150 × 3	$e \in R^{128}$	власна логіка

Далі наведено архітектури використаних нейромережевих моделей, а також опис наборів даних і процедур їх навчання, необхідних для реалізації запропонованого методу.

2.3.1 Архітектура та налаштування моделі детекції та класифікації об’єктів

Для задачі детекції та класифікації об’єктів у відеопотоці використано модель YOLOv11x з родини одноетапних детекторів YOLO (YouOnlyLookOnce), у яких локалізація та класифікація виконуються в межах одного прямого проходу мережею, без попереднього формування гіпотетичних областей-кандидатів. Такий підхід забезпечує суттєву обчислювальну перевагу порівняно з двоетапними детекторами та робить YOLO придатним для роботи з відеопотоком у реальному часі.

Архітектура YOLOv11 структурно складається з трьох частин: магістралі (backbone), шийки (neck) і голови (head). Магістраль побудована на блоках *C3k2* – розвитку блоків *C2f* моделі YOLOv8 з меншим ядром згортки у внутрішніх перетвореннях; вона послідовно знижує просторову розмірність та витягує ознаки зростаючого рівня абстракції з тензора $x \in R^{\{640 \times 640 \times 3\}}$. Шийка реалізує архітектуру *PathAggregationNetwork* (PANet) з двонаправленим обміном ознаками різного масштабу. Голова виконує якірно-вільне (anchor-free) передбачення прямокутників на трьох масштабах ознакових карт P_3, P_4, P_5, що відповідають об’єктам малого, середнього та великого розмірів. Загальну структурну схему моделі подано на рисунку 2.4.

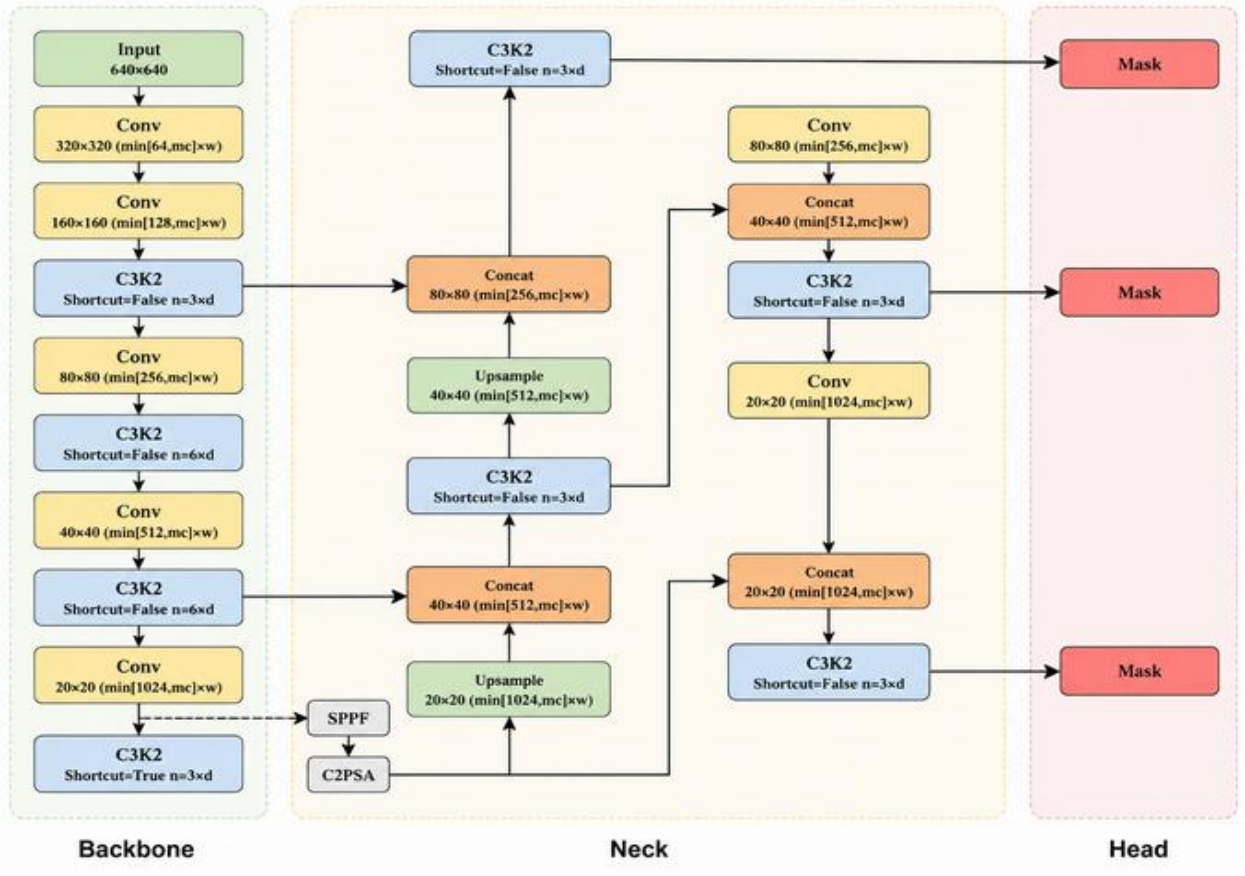


Рисунок 2.4 – Архітектура нейромережевої моделі YOLOv11 [34]

У межах запропонованого методу YOLOv11x використовується без донавчання, з ваговими коефіцієнтами, попередньо натренованими на датасеті COCO (80 класів). Конфігурація детектора: розмір входу 640 × 640, поріг впевненості τ_{conf} = 0,6, поріг впевненості для класу person τ_{conf}^{person} = 0,72, поріг *NMS* τ_{iou} = 0,45, активна підмножина класів обмежена переліком *ALLOWED_CLASSES* з дев'яти класів.

2.3.2 Архітектура та налаштування моделі ідентифікації та класифікації іменованих осіб

Для класифікації іменованих осіб реалізовано спеціалізовану згорткову нейронну мережу невеликої обчислювальної складності, що працює на прямокутних областях кадру, локалізованих детектором YOLOv11x. Вибір саме легкої CNN-архітектури обумовлено необхідністю повторного виконання класифікації багаторазово на кожному кадрі відеопотоку та обмеженими обчислювальними ресурсами цільової апаратної конфігурації [35].

Архітектура CNN-класифікатора приймає на вхід тензор $x \in R^{\{32 \times 32 \times 3\}}$, що відповідає області обличчя, отриманій з кадру за виразом:

$$x = resize\ (\ crop\ (\ I_t, b_t^k\ ),\ 32 \times 32\ )\ /\ 255, \tag{2.15}$$

де *crop* – операція виділення прямокутної області кадру за координатами b_t^k, *resize* – білінійне масштабування, ділення на 255 – нормалізація піксельних значень до діапазону [0, 1]. Послідовність шарів моделі подано у виразах:

$$h_1 = ReLU\ (\ Conv2D_{(32)}^{(3\times3)}\ (\ x\ )\ ) \tag{2.16}$$

$$h_2 = Dropout_{(0,25)}\ (\ MaxPool_{(2\times2)}\ (\ ReLU\ (\ Conv2D_{(64)}^{(3\times3)}\ (\ h_1\ )\ )\ )\ ) \tag{2.17}$$

$$h_3 = Dropout_{(0,25)}\ (\ MaxPool_{(2\times2)}\ (\ ReLU\ (\ Conv2D_{(128)}^{(3\times3)}\ (\ h_2\ )\ )\ )) \tag{2.18}$$

$$h_4 = Dropout_{(0,5)}\ (\ ReLU\ (\ Dense_{(256)}\ (\ Flatten\ (\ h_3\ )\ )\ )\ ) \tag{2.19}$$

$$y = softmax\ (\ Dense_{(L+1)}\ (\ h_4\ )\ ),\quad y \in \Delta^L \tag{2.20}$$

де Δ^L – симплекс імовірностей розмірності $L + 1$. Загальну структурну схему класифікатора подано на рисунку 2.5.

Розроблений CNN-класифікатор виконує роль спеціалізованого модуля розпізнавання іменованих осіб у межах областей, попередньо локалізованих детектором YOLOv11x.

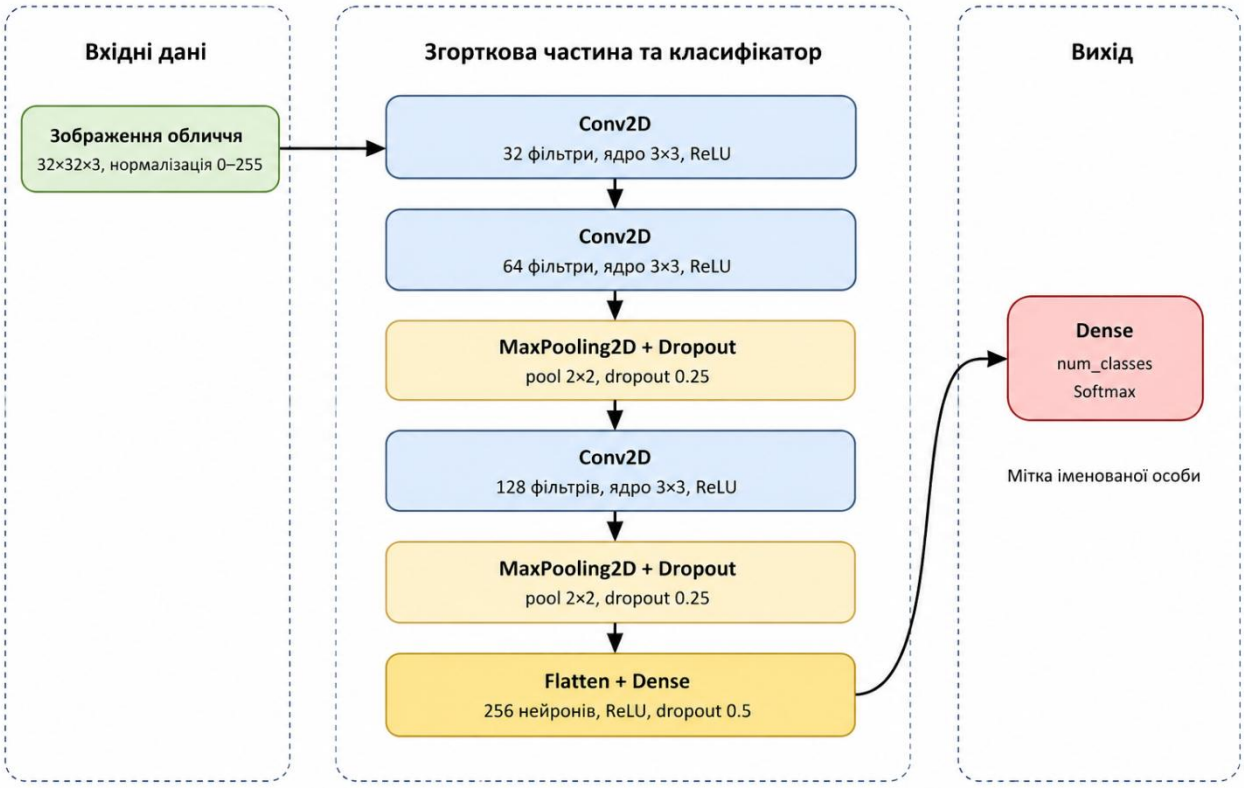


Рисунок 2.5 – Архітектура згорткової нейронної мережі для класифікації іменованих осіб

Запропонована архітектура поєднує кілька згорткових шарів для виділення просторових ознак, операції MaxPool для зменшення розмірності, Dropout для зниження ризику перенавчання та повнозв'язний блок із softmax-виходом для формування імовірнісного розподілу за класами. Використання вхідного зображення розміром 32 × 32 × 3 і відносно компактної структури моделі дає змогу зменшити обчислювальні витрати під час багаторазової класифікації фрагментів кадру у відеопотоці. Таким чином, обрана CNN-архітектура є компромісом між достатньою здатністю до розпізнавання візуальних ознак обличчя та вимогами до швидкодії системи в умовах обмежених ресурсів.

2.3.3 Формування та підготовка наборів даних для навчання моделей

Для оцінювання детектора об'єктів використовується датасет COCO та об'єднана валідаційна вибірка, сформована на основі публічних Kaggle-датасетів для дев'яти пріоритетних класів *ALLOWED_CLASSES*. Обсяг об'єднаної вибірки фіксується після первинної ручної верифікації розмітки. Перелік класів узгоджено з конфігурацією детектора (п. 2.3.1) та таблицею пріоритетів (п.2.1.4).

Для класифікатора іменованих осіб набір даних формується індивідуально для кожної зареєстрованої особи: підкаталог з відповідною назвою класу містить зображення обличчя у різних ракурсах, отримані з кадру кінцевим користувачем. Зображення попередньо обробляються відповідно до виразу (2.15): вирізаються за координатами обмежувального прямокутника, масштабуються до 32 × 32 та нормалізуються до діапазону [0, 1]. Мітки кодуються у форматі one-hotencoding.

2.3.4 Процедура навчання моделей

Навчання детектора YOLOv11x у межах роботи не виконується – використовуються попередньо натреновані на COCO ваги. Це рішення обґрунтоване достатнім покриттям класового простору *ALLOWED_CLASSES*

класами COCO та необхідністю забезпечити відтворюваність експерименту на типовій апаратній конфігурації.

Навчання CNN-класифікатора виконується у режимі стохастичного градієнтного спуску з оптимізатором Adam та категоріальною перехресно-ентропійною функцією втрат

$$L(y, y^*) = -\Sigma_{(l=1)}^{(L+1)} y_l^* \cdot \log(y_l), \qquad (2.21)$$

де y – вектор передбачених імовірностей (2.20), y^* – еталонний вектор у форматі one-hot. Процедура навчання включає поділ даних на навчальну та валідаційну підмножини, формування міток у форматі one-hotencoding, мінімізацію (2.21) методом Adam із пакетами фіксованого розміру $batch_{size}$ та кількістю епох *epochs*, моніторинг точності на валідаційній підмножині та критерій ранньої зупинки за відсутністю поліпшення метрики протягом фіксованої кількості епох. Конкретні значення $batch_{size}$ та *epochs* визначаються експериментально та розглядаються в розділі 3. Алгоритм навчання класифікатора схематично подано на рисунку 2.6.

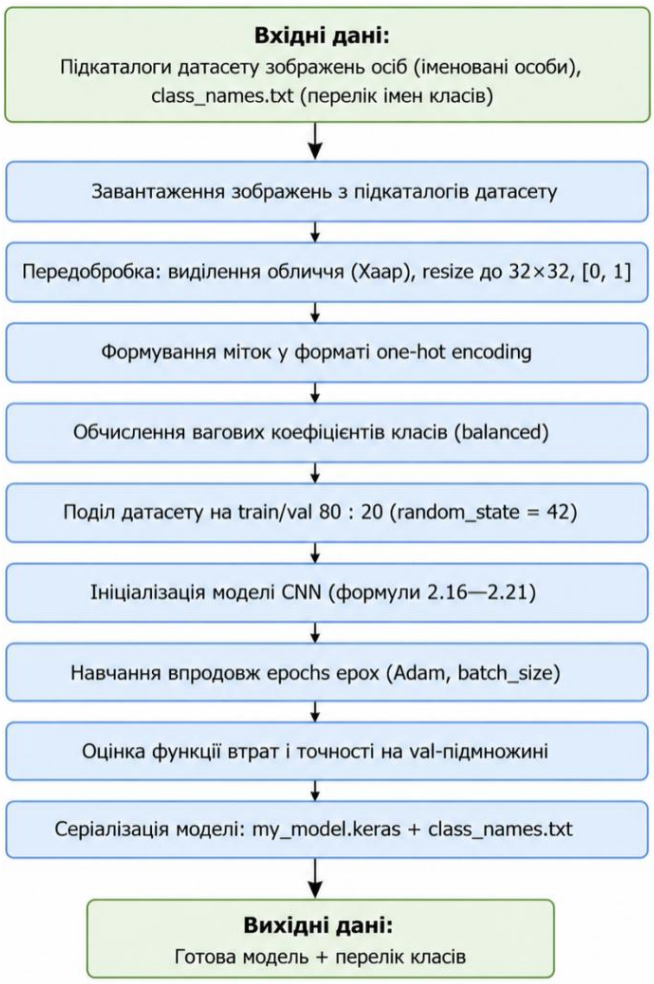


Рисунок 2.6 – Алгоритм навчання НМ для класифікації іменованих осіб

Отже, у межах роботи застосовується комбінований підхід, за якого детектор YOLOv11x використовується без додаткового навчання, що зменшує обчислювальні витрати та забезпечує відтворюваність експерименту. Основне навчання зосереджується на CNN-класифікаторі, який адаптується до цільової задачі шляхом мінімізації категоріальної перехресно-ентропійної функції втрат за допомогою оптимізатора Adam. Використання поділу даних на навчальну та валідаційну підмножини, one-hot кодування міток, контролю точності та механізму ранньої зупинки дає змогу організувати навчання класифікатора як контрольовану та відтворювану процедуру.

2.4 Метрики оцінювання ефективності методу

Оцінювання ефективності запропонованого методу формування аудіопотоку доповненої реальності за відеоданими має дати відповідь на два змістовно різні запитання. По-перше, наскільки якісно виконуються базові операції комп’ютерного зору, на яких ґрунтується метод, – детекція та класифікація об’єктів у кадрі, а також ідентифікація іменованих осіб. По-друге, наскільки результати цих операцій, інтегровані в єдиний аудіальний опис сцени, є корисними для кінцевого користувача з порушеннями зору. Перше запитання характеризує метод як технічну реалізацію, друге – як інструментальну систему підтримки сприйняття середовища.

Порівняльне оцінювання виконуватиметься з низкою базових конфігурацій. Для рівня комп’ютерного зору розглядатимуться інші представники сімейства YOLO (YOLOv5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n), що дозволить обґрунтувати вибір саме YOLOv11x як детектора у складі методу. Для аудіорівня базовою конфігурацією слугуватиме спрощений варіант методу без етапу семантичної пріоритизації (тобто з безпосереднім переказом результатів детекції) – таке порівняння дозволить кількісно довести внесок оператора Φ (вираз 2.13) у якість аудіопотоку. Сценарій експериментів, тестові послідовності та умови порівняння описуються у розділі 2.5; кількісні результати – у розділі 3.

Зведений перелік груп метрик з посиланнями на розділи, де наводяться їх формальні означення, подано у таблиці 2.2.

Таблиця 2.2 – Групи метрик для оцінювання ефективності методу

Рівень	Група метрик	Конкретні метрики	Підпункт
Комп'ютерний зір	Детекція об'єктів	IoU, Precision, Recall, F1, mAP@0.5, mAP@0.5:0.95	2.4.1
Комп'ютерний зір	Класифікація іменованих осіб	Accuracy, Top-1, Top-5, Verificationaccuracy	2.4.2
Аудіопотік	Часові показники	Latency, FPS	2.4.3
Аудіопотік	Семантичні показники	Coverage, Semanticcorrectness, Priorityaccuracy	2.4.3
Аудіопотік	Суб'єктивні показники	MOS (MeanOpinionScore)	2.4.3

Логіка доведення ефективності методу у дослідженні така: 1) на рівні комп'ютерного зору отримати значення метрик з підрозділу 2.4.1 для YOLOv11x на тестовій вибірці COCO128 та значення метрик з підрозділу 2.4.2 для CNN-класифікатора на тестовій підмножині облич, узгоджені з базовими конфігураціями; 2) на рівні аудіопотоку отримати значення метрик з підрозділу 2.4.3 на тестових сценаріях, що моделюють реальні випадки практичного використання; 3) встановити, що отримані значення задовольняють орієнтири, визначені у п. 2.4.3 (зокрема $Latency_{avg}<0,5c, Latency_{avg}< 0,8c$ *і* $FPS \geq 30$), і що внесок етапу семантичної пріоритизації у Coverage та Priorityaccuracy є позитивним порівняно з базовим варіантом методу без оператора Φ.

2.4.1 Метрики детекції та класифікації

Даний підрозділ розкриває першу групу метрик першого рівня, що характеризує якість роботи детектора об'єктів у складі запропонованого методу. Усі показники обчислюються на тестовій вибірці зі статичною покадровою розміткою, у якій для кожного значущого об'єкта вказано еталонний

обмежувальний прямокутник і клас. Логіка зіставлення передбачень з еталонами єдина: для кожного передбаченого прямокутника відшукується еталонний прямокутник того самого класу, що має з ним найбільший показник *IoU*; якщо це значення не нижче порога τ_{IoU}, передбачення класифікується як істинно позитивне (*TP*), інакше – як хибнопозитивне (*FP*). Еталонний прямокутник, для якого жодне з передбачень не задовольнило цю умову, дає одне хибнонегативне (*FN*). У такий спосіб для кожного кадру і для кожного класу окремо формуються лічильники *TP, FP, FN*, з яких далі обчислюються усі похідні показники.

Основною геометричною мірою узгодженості передбаченого та еталонного обмежувальних прямокутників є показник *IoU*:

$$IoU(b', b) = |b' \cap b| / |b' \cup b|, \qquad (2.22)$$

де b' – передбачений прямокутник, b – еталонний прямокутник, $|\cdot|$ – площа. Значення *IoU* лежить у межах [0, 1]: 0 означає відсутність перетину, 1 – ідеальний збіг. Для двох прямокутників, заданих координатами *(x1, y1, x2, y2)*, перетин обчислюється як прямокутник з координатами $(\max(x1^p, x1^g), \max(y1^p, y1^g), \min(x2^p, x2^g), \min(y2^p, y2^g))$, а об’єднання – як $|b'| + |b| - |b' \cap b|$.

Приклад. Нехай передбачений прямокутник має площу 200 px², еталонний – 180 px², а площа їх перетину дорівнює 120 px². Тоді площа об’єднання – 200 + 180 − 120 = 260 px², і *IoU* = 120 / 260 ≈ 0,46. За порога τ_{IoU} = 0,5 таке передбачення буде класифіковано як *FP*, а за порога τ_{IoU}= 0,4 – як *TP*.

Поріг τ_{IoU} є гіперпараметром процедури оцінювання, який задає жорсткість вимоги до точності локалізації. У роботі використовуються два значення порога: τ_{IoU} = 0,5 (типове для детекторів загального призначення) та сітка значень {0,5; 0,55; …; 0,95} – для обчислення *mAP@0.5:0.95* (вираз 2.29 нижче).

На основі підрахунку *TP, FP* та *FN* означаються показники точності (*Precision*) і повноти (*Recall*):

$$Precision = TP / (TP + FP) \qquad (2.23)$$

$$Recall = TP / (TP + FN) \qquad (2.24)$$

Precision характеризує частку правильних спрацьовувань серед усіх повернутих моделлю – тобто наскільки можна довіряти позитивній відповіді

детектора. *Recall* характеризує частку справжніх об’єктів, яких модель змогла знайти, – тобто наскільки повно детектор охоплює сцену. Для прикладної задачі формування аудіопотоку обидва показники однаково важливі: низька *Precision* призводить до надлишкових повідомлень про неіснуючі об’єкти, що дезорієнтує користувача; низький *Recall* призводить до пропусків значущих об’єктів і втрати інформативності аудіопотоку [36].

Числовий приклад. Нехай для класу person у тестовій вибірці налічується 200 еталонних об’єктів. Детектор повернув 180 передбачень класу person, з яких 150 узгоджені з еталонами за порогом $\tau_{IoU} = 0{,}5$ (*TP*), а 30 – ні (*FP*). Не виявлених еталонів – $200 - 150 = 50$ (*FN*). Тоді *Precision* = 150 / 180 ≈ 0,833, *Recall*= 150 / 200 = 0,750, $F_1 \approx 0{,}789$. Подальша зміна порога впевненості моделі переноситиме об’єкти між *TP*, *FP* та *FN* і змінюватиме значення *Precision* і *Recall* у протилежних напрямах, що і дає основу для побудови *PR*-кривої (вираз 2.26 нижче).

Збалансованою комбінацією *Precision* і *Recall*є F_1-міра – гармонічне середнє цих показників:

$$F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall) . \qquad (2.25)$$

F_1 досягає максимуму у разі рівності *Precision* і *Recall* і штрафує конфігурації, у яких один із двох показників суттєво перевищує інший. Це робить F_1 зручною для покласового аналізу: клас, для якого один з показників наближається до нуля, отримує низьке значення F_1 незалежно від іншого показника.

Інтегральною мірою якості детектора є середня усереднена точність *mAP*. Її обчислення складається з двох кроків: для кожного класу обчислюється середня точність *AP* – площа під кривою *Precision–Recall* – а потім значення *AP* усереднюються за всіма класами. Крива *Precision–Recall* будується шляхом сортування передбачень класу за значенням confidence у порядку спадання та послідовного обчислення (*Precision, Recall*) для кожного конкретного порогового рівня впевненості. Інтегральна площа під цією кривою для класу c означається як:

$$AP_c = \int_0^1 Precision_c(Recall)\, d\, Recall. \quad (2.26)$$

На практиці інтеграл обчислюється чисельно: у класичному варіанті *PASCAL VOC* – через інтерполяцію *PR*-кривої в 11 точках $Recall \in \{0; 0{,}1; \ldots; 1\}$; у варіанті COCO, що відповідає реалізації Ultralytics, – через усереднення *Precision* у 101 точці $Recall \in \{0; 0{,}01; \ldots; 1\}$. Усереднення значень *AP* за множиною всіх класів C дає інтегральний показник *mAP*:

$$mAP = (1 / |C|) \cdot \Sigma_{(c \in C)} AP_c. \quad (2.27)$$

У роботі обчислюватимуться два варіанти *mAP*, що відрізняються способом задання порога *IoU*. *mAP@0.5* обчислюється за фіксованого порога τ_{IoU}= 0,5: *TP*, *FP*, *FN* для кожного класу формуються один раз за цим порогом, далі будується *PR*-крива і обчислюється AP_c, після чого результати усереднюються за класами. У такому режимі достатньо, щоб передбачений прямокутник перетинався з еталонним хоча б наполовину; це робить показник «м'яким» до локалізаційних похибок і характеризує радше якість класифікації, ніж точність локалізації:

$$mAP@0.5 = mAP \text{ при } \tau_{IoU} = 0{,}5. \quad (2.28)$$

Значення *mAP@0.5:0.95* обчислюється як середнє значення *mAP* за дискретною сіткою порогів *IoU* від 0,5 до 0,95 із кроком 0,05:

$$mAP@0.5:0.95 = (1 / 10) \cdot \Sigma_{(\tau \in \{0{,}5;\ 0{,}55;\ \ldots;\ 0{,}95\})} mAP(\tau) \quad (2.29)$$

Цей показник є «жорсткішим»: він знижується, якщо передбачені прямокутники локалізовані недостатньо точно, навіть за умови правильної класифікації. *mAP@0.5:0.95* краще відображає здатність детектора видавати точні координати об'єкта і саме він є обов'язковим показником у бенчмарку COCO та у бібліотеці Ultralytics.

Порівняння двох варіантів є показовим. Чисельний приклад: модель, що відрізняється високою впевненістю класифікації, але слабкою локалізацією (значення *IoU* здебільшого у діапазоні 0,5–0,6), матиме високий *mAP@0.5* (0,80) і відчутно нижчий *mAP@0.5:0.95* (0,45). Натомість модель з точною локалізацією (*IoU*здебільшого у діапазоні 0,7–0,9) утримує високі значення обох показників: *mAP@0.5* близько 0,80 і *mAP@0.5:0.95* близько 0,65. Різниця (*mAP@0.5 − mAP@0.5:0.95*) у такий спосіб характеризує локалізаційну точність детектора: чим вона менша, тим точніші координати прямокутників. Для

запропонованого методу формування аудіопотоку доповненої реальності пріоритетною є висока точність розпізнавання об’єктів пріоритетних класів (person, car, motorcycle, bicycle, bus, truck, trafficlight, stopsign, knife) при припустимих обчислювальних витратах. Орієнтовні цільові значення метрик детекції та класифікації подано у таблиці 2.3.

Таблиця 2.3 – Орієнтовні цільові значення метрик детекції та класифікації

Показник	**Формула**	**Цільове значення**	**Зміст**
IoU	(2.22)	≥ 0,50	Геометрична узгодженість прямокутників
Precision	(2.23)	≥ 0,70	Довіра до позитивної відповіді детектора
Recall	(2.24)	≥ 0,60	Повнота охоплення об’єктів сцени
F_1	(2.25)	≥ 0,65	Збалансованість Precision і Recall
mAP@0.5	(2.28)	≥ 0,70	Інтегральна якість у режимі грубої локалізації
mAP@0.5:0.95	(2.29)	≥ 0,50	Інтегральна якість з урахуванням локалізації

Перевищення цих значень для обраної моделі YOLOv11x порівняно з конфігураціями YOLOv5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n дозволить обґрунтувати її вибір як нейромережевої моделі для детекції та класифікації об’єктів у складі методу. Покласові значення *Precision, Recall,mAP@0.5* та *mAP@0.5:0.95* для YOLOv11x на тестовій вибірці COCO128 наводяться у розділі 3.2.

2.4.2 Метрики розпізнавання іменованих осіб

Розділ розкриває другу групу метрик першого рівня, що характеризує якість роботи CNN-класифікатора іменованих осіб. Оцінювання виконується на тестовій вибірці зображень обличчя з відомими еталонними мітками;

кожне зображення обличчя вирізається з кадру за обмежувальним прямокутником класу «person», повертеним детектором, і нормалізується відповідно до виразу (2.15).

Базовим показником якості класифікатора є точність розпізнавання, що дорівнює частці правильно класифікованих тестових зображень:

$$Accuracy = (1 / N) \cdot \Sigma_{n=1}^{N} \mathbb{1} [\hat{y}_n = y_n], \quad (2.30)$$

де N – кількість тестових зображень, $\hat{y}_n$ – передбачена мітка, y_n – еталонна мітка, $\mathbb{1}[\cdot]$ – індикаторна функція. *Accuracy* є достатньою як інтегральна оцінка лише для збалансованих за класами вибірок; у випадку дисбалансу між класами (перевагою класу other) точність може приховувати погану якість розпізнавання окремих іменованих осіб, тому її доповнюють *Top-K* та покласовими*Precision/Recall/*F_1 (формули 2.23–2.25 з підрозділу 2.4.1).

Якщо результатом класифікатора є розподіл імовірностей за класами (softmax-вихід), для більш повного оцінювання використовуються *Top-K* показники:

$$Top\text{-}K = (1 / N) \cdot \Sigma_{n=1}^{N} \mathbb{1} [y_n \in Top\text{-}K (y(x_n))], \quad (2.31)$$

де *TopK(·)* – множина з K класів, що мають найбільші значення ймовірності. У роботі обчислюватимуться *Top-1* (фактично збігається з *Accuracy* при N класах ≥ 5) та *Top-5*, що характеризує, як часто еталонна особа потрапляє у п’ять найбільш імовірних варіантів класифікатора. *Top-5* у задачі з невеликою кількістю іменованих осіб ($L \approx 20$) є м’якшим, ніж *Top-1*, і дає змогу зафіксувати випадки «класифікатор зрозумів правильну особу, але не виставив її на перше місце».

Для пар зображень обличчя у режимі верифікації, що відповідає пороговому правилу (2.12), використовується показник *Verificationaccuracy*. Він оцінює якість прийняття рішення «свій/чужий» на тестових парах, у яких відомо, чи обидва зображення належать одній особі. Формально:

$$Verificationaccuracy = (TP_v + TN_v) / (TP_v + TN_v + FP_v + FN_v), \quad (2.32)$$

де матриця помилок (TP_v, TN_v, FP_v, FN_v) обчислюється на тестових парах ембедингів обличчя: пара (i, j) вважається позитивною, якщо для неї $d(e_i, e_j) < \tau$ і

обидва обличчя справді належать одній і тій самій зареєстрованій особі. Значення τ узгоджене з порогом порівняння у застосунку (*TOLERANCE* = 0,48).

Орієнтовні цільові значення для запропонованого методу: *Accuracy≥ 0,95, Top-1 ≥ 0,95, Top-5 ≥ 0,99*, *Verification accuracy* ≥ 0,95. Перевищення цих значень одночасно з високою *Precision* класу «other» є необхідною умовою надійної ідентифікації іменованих осіб в інтегральному конвеєрі методу.

2.4.3 Метрики якості формування аудіопотоку доповненої реальності

Розділ розкриває другий, інтегральний рівень метрик, який характеризує ефективність методу як цілісної системи формування аудіопотоку, а не лише якість окремих нейромережевих компонентів. Цей рівень є визначальним для обґрунтування корисності методу для користувача з порушеннями зору, оскільки високі значення метрик комп’ютерного зору з підрозділів 2.4.1, 2.4.2 не гарантують, що система формує своєчасні, повні та змістовно правильні аудіоповідомлення в умовах реального відеопотоку [37].

Часова характеристика – латентність – означається як інтервал між моментом надходження j-го кадру $t_j^{(frame)}$ та моментом запуску синтезу мовлення для відповідного повідомлення τ_j [38]:

$$Latency_j = \tau_j - t_j^{(frame)} \tag{2.33}$$

Середня латентність та її 95-перцентиль на тестовій послідовності з M повідомлень обчислюються як:

$$Latency_{avg} = (1 / M) \cdot \Sigma_{j=1}^{M} Latency_j \tag{2.34}$$

$$Latency_{p95} = Quantile_{0,95}(Latency_j) \tag{2.35}$$

$Latency_{avg}$ характеризує середній досвід користувача, а $Latency_{p95}$ – «найгірший» практично значущий випадок: 0,95 повідомлень формуються не довше за це значення. Для інтерактивних систем доповненої реальності прийнятою верхньою межею латентності вважається 1 с; цільовою для методу обрано $Latency_{\text{avg}} \leq 0{,}5$ с і $Latency_{\{p95\}} \leq 0{,}8$ с.

Швидкодія обробки відеопотоку оцінюється показником *FPS* (FramesPerSecond), що визначається як середня кількість кадрів, оброблених методом за одну секунду:

$$FPS = N_{frames} / T_{total}, \quad (2.36)$$

де N_{frames} – кількість оброблених кадрів, T_{total} – загальна тривалість обробки. Показник *FPS* включений у систему метрик тому, що для застосунку, орієнтованого на роботу з живим відеопотоком, недостатньо лише низької затримки одного повідомлення: метод має «встигати» обробляти послідовність кадрів у режимі, наближеному до реального часу. Цільовим є $FPS \geq 30$ кадрів/с на типовій апаратній конфігурації, що відповідає налаштованому інтервалу 30 мс таймера головного циклу обробки [39].

Coverage характеризує повноту охоплення сцени аудіопотоком – тобто частку значущих об’єктів сцени, які потрапили до повідомлень:

$$Coverage = | A_{obj} | / | S_{obj}^{*} |, \quad (2.37)$$

де A_{obj} – множина об’єктів, озвучених у потоці A; S_{obj}^{*} – еталонна множина значущих об’єктів сцени, отримана покадровою розміткою. *Coverage* обмежено зверху роботою детектора: пропуски детектора неминуче знижують *Coverage*. Цільовим є $Coverage \geq 0{,}90$.

Семантична коректність (*Semantic correctness*) оцінює, яка частка сформованих повідомлень правильно описує сцену – тобто і текстова мітка, і словесна ознака просторового положення відповідають реальному змісту кадру:

$$Semanticcorrectness = (1 / M) \cdot \Sigma_{j=1}^{M} \mathbb{1} [(l_j, z_j) = (l_j^{*}, z_j^{*})], \quad (2.38)$$

де (l_j, z_j) – пара «мітка + позиція», сформована методом; (l_j^{*}, z_j^{*}) – еталонна пара. Цей показник найбільшою мірою відображає корисність повідомлень для кінцевого користувача, оскільки об’єднує і коректність розпізнавання класу, і коректність опису просторової конфігурації сцени. Цільовим є *Semantic correctness* $\geq 0{,}93$.

Точність пріоритизації (*Priority accuracy*) характеризує, наскільки фактичний порядок озвучення повідомлень узгоджений із еталонним пріоритетним порядком, заданим таблицею пріоритетів (п. 2.1.4). Як міра

неузгодженості двох перестановок використовується нормована відстань Кендалла-Тау [40]:

$$Priority\ accuracy = 1 - (DKT (ord_A, ord_*) / ord_{max}), \quad (2.39)$$

де ord_A – фактична послідовність озвучення, ord_* – еталонна, $DKT(\cdot, \cdot)$ – кількість інверсій (пар повідомлень, що поміняли місцями), $ord_{max} = M(M - 1)/2$ – максимально можлива кількість інверсій для перестановки довжини M. Значення 1 означає повний збіг порядків, 0 – повну невідповідність. Цільовим є *Priority accuracy* $\geq 0{,}95$.

Суб’єктивна якість сприйняття аудіопотоку оцінюється усередненим показником *MOS* (MeanOpinionScore), що обчислюється за оцінками R опитаних респондентів [41]:

$$MOS = (1 / R) \cdot \Sigma_{r=1}^{R} q_r, \quad q_r \in \{1, 2, 3, 4, 5\} \quad (2.40)$$

де q_r – оцінка r-го респондента за п’ятибальною шкалою (1 – повністю неприйнятно, 5 – повністю задовільно). *MOS* наведено як перспективний показник для подальших досліджень із залученням цільової групи користувачів із порушеннями зору; у межах цієї роботизначенням є *MOS* $\geq$ 4,0 на тестовій групі.

Зведення цільових значень метрик аудіорівня подано у таблиці 2.4.

Таблиця 2.4 – Орієнтовні цільові значення метрик якості формування аудіопотоку

Показник	**Формула**	**Цільове значення**	**Зміст**
$Latency_{avg}$	(2.34)	$\leq 0{,}5$ с	Своєчасність повідомлень у середньому
$Latency_{\{p95\}}$	(2.35)	$\leq 0{,}8$ с	Своєчасність 0,95 повідомлень
FPS	(2.36)	≥ 30	Швидкодія обробки кадрів
Coverage	(2.37)	$\geq 0{,}90$	Повнота охоплення значущих об’єктів
Semanticcorrectness	(2.38)	$\geq 0{,}93$	Коректність мітки та положення
Priorityaccuracy	(2.39)	$\geq 0{,}95$	Узгодженість порядку озвучення
MOS	(2.40)	$\geq 4{,}0$ / 5	Суб’єктивна якість сприйняття (допоміжний)

Цей блок метрик є визначальним «унікальним» рівнем оцінювання запропонованого методу: на відміну від загальноприйнятих метрик комп’ютерного зору з п. 2.4.1 і 2.4.2, він прив’язаний саме до задачі формування аудіопотоку доповненої реальності і одночасно відображає часовий, семантичний та суб’єктивний аспекти якості повідомлень. Сценарій експериментального обчислення цих показників наведено у п. 2.5, а кількісні результати на тестових послідовностях – у п.3.4.

2.5 Сценарій експериментального дослідження

Метою експериментального дослідження є оцінювання ефективності запропонованого методу на двох рівнях: компонентному (детекція об’єктів та класифікація іменованих осіб) та системному (інтегральне формування аудіопотоку), шляхом застосування відповідних метрик якості (2.22–2.40) у стандартизованих тестових сценаріях вуличного, кімнатного та комбінованого типів.

Гіпотеза дослідження полягає в тому, що багаторівнева інтеграція нейромережевої детекції об’єктів, класифікації іменованих осіб та семантично керованого механізму формування аудіопотоку забезпечує підвищення інформативності, повноти та семантичної коректності аудіального представлення сцени, а також зниження затримки його формування, порівняно з конфігурацією, що використовує лише детекцію об’єктів без подальшої семантичної інтерпретації та ідентифікації осіб.

Перший рівень експериментів передбачає оцінювання якості детектора YOLOv11x та CNN-класифікатора за метриками (2.22)–(2.32). Порівняння виконується з базовими конфігураціями YOLOv5n, YOLOv5x, YOLOv8n,

YOLOv8x, YOLOv11n. Другий рівень передбачає оцінювання інтегральної якості методу за метриками (2.33)–(2.40) на тестовій вибірці відео.

Тестові сценарії дібрані так, щоб охопити основні випадки практичного використання:

– вулична сцена з транспортом: 3–5 припаркованих або рухомих автомобілів і пішоходи;

– кімнатна сцена з іменованою особою: приміщення зі зниженим освітленням, у центрі – обличчя зареєстрованого користувача;

– комбінована сцена: перехід з кімнати у двір з іменованою особою поруч з автомобілем.

Апаратна конфігурація: центральний процесор IntelCore i5-11400H, 16 ГБ оперативної пам’яті DDR4 (3200 MT/c), графічний прискорювач NVIDIA GeForce RTX 3050 Laptop GPU. Програмне середовище – Python з пакетами tensorflow, ultralytics, opencv-python, face_recognition, pyttsx3, PyQt5, pandas, numpy. Умови тестування уніфіковано: однакові тестові вибірки, фіксовані правила поділу даних, узгоджені критерії формування аудіоповідомлень. Результати оцінювання та їх інтерпретація подаються у розділі 3.

2.6 Висновки до розділу 2

У другому розділі формалізовано задачу формування аудіопотоку доповненої реальності за відеоданими. Введено основні позначення $(V, I_t, O_t, o_t^k, S_t, s_t^k, A, a_j)$, описано загальну композицію перетворень $V \to O \to S \to A$ та її чотири підзадачі (п.п. 2.1.1–2.1.4). Для кожної підзадачі математичної моделі, крім формалізації вхідних та вихідних даних, побудовано математичний псевдокод (алгоритми 2.1–2.3).

У розділі 2.2 розкрито метод формування аудіопотоку як композицію чотирьох послідовних етапів (попередня обробка, детекція та класифікація, ідентифікація осіб, формування аудіоопису) та подано загальну архітектуру програмного забезпечення інтелектуальної системи. У розділі 2.3 описано архітектуру та налаштування двох нейромережевих компонентів – детектора YOLOv11x та CNN-класифікатора іменованих осіб – і процедуру їх навчання.

У розділі 2.4 побудовано дворівневу систему метрик: метрики комп’ютерного зору (підрозділи 2.4.1, 2.4.2) та метрики якості формування аудіопотоку (п. 2.4.3). Усі показники подано формулами (2.22)–(2.40) із коротким змістовним поясненням.

У п. 2.5 визначено сценарій експериментального дослідження методу, перелік базових конфігурацій порівняння, тестові сценарії та апаратно-програмне середовище. Сформована у розділі теоретико-методична основа становить опору для подальшого програмного втілення методу та експериментального дослідження, результати якого подаються у третьому розділі.

Розділ 3 Експериментальне дослідження методу та застосування інтелектуальної системи

3.1 Опис інтелектуальної системи для формування аудіопотоку доповненої реальності за відеоданими

Предметною областю інтелектуальної системи для формування аудіопотоку доповненої реальності за відеоданими є інтелектуальні асистивні технології для людей із порушеннями зору, зокрема системи комп'ютерного зору, які аналізують відеодані та передають користувачеві інформацію про навколишню сцену в аудіальній формі. Такий підхід поєднує методи детекції об'єктів, розпізнавання облич, семантичної інтерпретації сцени та синтезу мовлення.

Система призначена для використання в ситуаціях, де користувачеві необхідно отримати короткий аудіальний опис важливих об'єктів навколо нього. До таких ситуацій належать переміщення у приміщеннях, орієнтування в навчальному або робочому середовищі, виявлення людей поруч, розпізнавання знайомих осіб, а також отримання попереджень про потенційно важливі або небезпечні об'єкти в полі зору камери.

На відміну від звичайного відеоаналізу, результат роботи системи не обмежується візуальним відображенням рамок навколо об'єктів. Основним результатом є аудіопотік, сформований із урахуванням класу об'єкта, його просторового положення, рівня впевненості моделі та пріоритету озвучення.

Кінцевим користувачем інтелектуальної системи є людина з повною або частковою втратою зору, яка потребує додаткової інформації про навколишній простір. Для такого користувача важливими є не всі об'єкти сцени, а лише ті, що мають практичну значущість у конкретний момент часу: люди, транспортні засоби, перешкоди, меблі, двері, побутові предмети або інші об'єкти з множини допустимих класів *ALLOWED_CLASSES*.

Типовий сценарій використання системи передбачає запуск графічного застосунку, вибір джерела відеоданих і подальше отримання аудіальних повідомлень про об'єкти в полі зору камери. Користувач може використовувати

вебкамеру або заздалегідь записаний відеофайл. Після запуску система послідовно отримує кадри I_t, виконує їх попередню обробку, подає кадри на вхід моделі детекції, формує множину об’єктів *Ot*, уточнює мітки для об’єктів класу person і перетворює результати аналізу на семантичне представлення сцени S_t.

Наприклад, якщо в кадрі виявлено людину зліва від користувача, система формує повідомлення на зразок person on your left. Якщо обличчя цієї людини збігається з одним із еталонних ембедингів у базі, загальна мітка person може бути замінена на ім’я іменованої особи. У такому випадку аудіоповідомлення стає більш інформативним, оскільки користувач отримує не лише відомості про наявність людини, а й інформацію про її ідентичність.

3.1.1 Проєктування та програмна реалізація

Інтелектуальна система реалізує метод формування аудіопотоку доповненої реальності за відеоданими, формально описаний у другому розділі (вирази 2.1–2.14). Програмну реалізацію системи здійснено як графічний застосунок мовою Python із використанням бібліотек PyQt5 [42] для побудови інтерфейсу користувача, OpenCV [43] – для отримання та базової обробки кадрів, Ultralytics YOLO [44] і PyTorch [45] – для детекції об’єктів, face_recognition [46] і Keras [47] – для розпізнавання та класифікації іменованих осіб, а також pyttsx3 [48] і QtMultimedia [49] – для синтезу мовлення та відтворення амбієнтного звуку.

Структурно застосунок поділено на чотири взаємопов’язані функціональні модулі (рисунок 3.1). Перший модуль відповідає за керування джерелом відео та інтерфейсом користувача. Другий модуль виконує детекцію та класифікацію об’єктів у кадрі на основі ваг YOLO. Третій модуль виконує ідентифікацію іменованих осіб для об’єктів класу person. Четвертий модуль формує аудіопотік доповненої реальності, що включає мовні повідомлення про значущі об’єкти та фонове звукове супроводження.

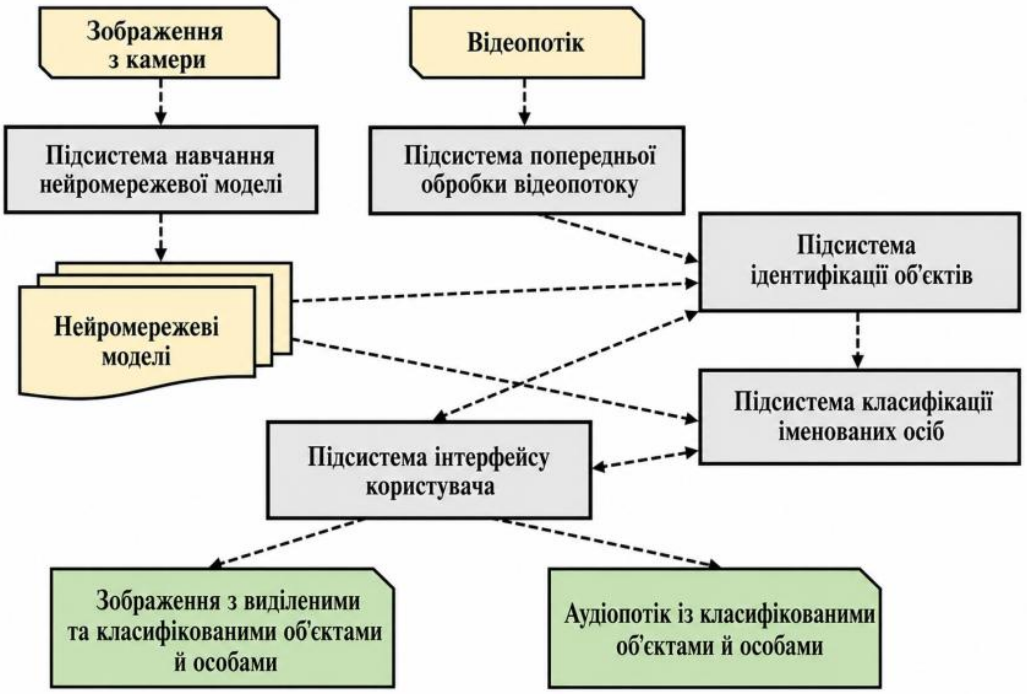


Рисунок 3.1 – Функціональні модулі

Модуль користувацького інтерфейсу побудовано як головне вікно класу *DetectorGUI*, що містить область попереднього перегляду відеокадру, текстове поле для виведення журналу детекцій, вибір індексу камери та керівні кнопки. Користувач може запустити роботу з камерою через елемент керування «StartCamera», обробку файлу через «OpenVideo …», зупинити обробку через «Stop», додати до бази облич нову іменовану особу через «AddPerson …», а також увімкнути або вимкнути амбієнтне звукове супроводження через прапорець «Ambientsound». Послідовність обробки кадрів реалізовано через QTimer з інтервалом 30 мс, що відповідає цільовій частоті надходження кадрів приблизно 33 за секунду.

Модуль детекції та класифікації об’єктів використовує ваги моделі YOLOv11x, завантажені через інтерфейс бібліотеки Ultralytics. Перед запуском обчислень модель переводиться у режим оцінювання та, за наявності CUDA, у режим напівточних обчислень. Для кожного кадру вхідне зображення переводиться у формат RGB, після чого виконується інференс із розміром вхідного боку 640 пікселів. Результат інференсу фільтрується за класовим простором, заданим списком *ALLOWED_CLASSES*, що визначає множину семантичних міток, релевантних для прикладної задачі підтримки сприйняття

навколишнього середовища. Перелік відповідних класів і пов’язаних із ними пріоритетів наведено у таблиці 3.1.

Таблиця 3.1 – Класи об’єктів, які озвучуються інтелектуальною системою, та їхні пріоритети

Семантична мітка	**Категорія**	**Пріоритет озвучення**	**Особливості**
(іменована особа)	людина	0	Найвищий пріоритет; формується через face_recognition + CNN
person	людина	1	Нерозпізнана особа; озвучується як загальна мітка
car	транспорт	2	Активує амбієнтний звук transport.wav
motorcycle	транспорт	3	Активує амбієнтний звук transport.wav
bus	транспорт	4	Активує амбієнтний звук transport.wav
bicycle	транспорт	5	Активує амбієнтний звук transport.wav
knife	небезпечний предмет	6	Озвучується як можлива загроза
scissors	небезпечний предмет	7	Озвучується як можлива загроза
fork	побутовий предмет	8	Озвучується як побутовий об’єкт

Підмодуль розпізнавання іменованих осіб активується для тих результатів детекції, що відповідають класовому індексу person. Для відповідного обмежувального прямокутника на кадрі формується область інтересу із невеликим запасом і передається у функцію *face_recognition.face_encodings* для побудови 128-вимірного вектора подання обличчя. Отриманий вектор порівнюється з усіма еталонними векторами, попередньо завантаженими з локальної бази *known_faces.pkl*. Рішення про збіг приймається на основі мінімальної евклідової відстані між поточним та еталонним векторами; для прийняття позитивного рішення вимагається, щоб ця відстань не перевищувала наперед заданого порогу *TOLERANCE* = 0,48. Якщо рішення позитивне, семантична мітка person замінюється на ім’я відповідної особи; інакше мітка залишається загальною. Для зменшення обчислювального навантаження процедура побудови ембедингу виконується не для кожного кадру, а раз на

$EVERY_N$ = 4 кадри, з кешуванням відповідності між обмежувальним прямокутником і присвоєним іменем.

Окремим компонентом системи є CNN-класифікатор іменованих осіб, сформований у скрипті train.py. Він має структуру, описану формулами (2.15)–(2.21) другого розділу, і використовує вхідне зображення розміром 32 × 32. Класифікатор зберігається у файл my_model.keras та забезпечує альтернативний шлях ідентифікації іменованої особи в умовах, коли векторне подання облич недостатньо стабільне (низьке освітлення, нестандартний ракурс). У програмному застосунку результат класифікатора інтегрується з порівнянням ембедингівface_recognition: остаточне рішення формується через мажоритарну логіку між двома гілками ідентифікації.

Формування аудіопотоку реалізовано через два паралельні підмодулі. Перший підмодуль – клас *VoiceAnnouncer*, який забезпечує озвучення семантичних міток через pyttsx3. Чергу повідомлень побудовано на основі мінімальної купи, упорядкованої за пріоритетом, заданим у таблиці 3.1, із вторинною сортувальною ознакою у вигляді мітки часу появи події. Перед озвученням кожне повідомлення проходить перевірку на актуальність: повідомлення, що перебуває в черзі довше за $STALE_{\mathrm{T}}$ = 2,0 с, відкидається, що відповідає логіці часової актуальності для систем реального часу. Зміст повідомлення будується за шаблоном «{семантична мітка} onyour {left|ahead|right}», де словесна ознака положення обчислюється за відношенням центра обмежувального прямокутника до ширини кадру.

Другий підмодуль – клас *AmbientSoundManager*, що відповідає за фонове звукове супроводження присутніх у сцені транспортних засобів. Підмодуль циклічно відтворює файл sounds/transport.wav, а його гучність динамічно регулюється у залежності від відносної площі найбільшого транспортного об’єкта у кадрі: для відносної площі менше за *MIN_REL_AREA* = 0,008 фоновий звук вимикається, для більшої – масштабується множником *SCALE* = 5,0 з обмеженням верхньою межею MAX_{VOL} = 0,4. Така схема дає змогу користувачеві опосередковано оцінювати наближення транспортного засобу за інтенсивністю фонового звуку, не перевантажуючи мовний канал повідомлень.

Загальна послідовність обробки одного кадру у застосунку відповідає композиції $V \rightarrow O \rightarrow S \rightarrow A$, формалізованій у виразі (2.1) другого розділу. У

термінах коду main.py ця послідовність реалізована методом _next() класу *DetectorGUI*: отримання кадру з cv2.*VideoCapture*; інференс моделі YOLO; фільтрація за класовим простором; для кожного об’єкта – обчислення площі обмежувального прямокутника, словесної ознаки положення та пріоритету; для класу person – виклик підмодуля розпізнавання обличчя; формування пари (ключ, текст) і додавання її до черги *VoiceAnnouncer*; оновлення гучності *AmbientSoundManager*; візуалізація детекцій у вікні попереднього перегляду та запис відповідного рядка журналу. Завдяки тому що озвучення виконується в окремому потоці виконання, основний цикл обробки кадру не блокується.

Файлова структура проєкту узгоджена із наведеною композицією та подана у таблиці 3.2. Усі обчислювальні модулі винесено у відповідні підпапки, що дає змогу запускати окремо інтерактивний застосунок, навчання класифікатора та оцінювання детектора.

Таблиця 3.2 – Структурні складові програмної реалізації

Складова	**Шлях у проєкті**	**Призначення**
Інтерактивний застосунок	app/main.py	Графічний застосунок PyQt5 з детекцією YOLO, ідентифікацією осіб і формуванням аудіопотоку
Звукові ресурси застосунку	app/sounds/	Звукові ефекти для інтерфейсу та амбієнтний шум транспорту
Підготовка датасетів	datasets/	Архіви наборів даних, що використовуються під час навчання та оцінювання моделей
Оцінювання детектора	metrics/evaluate.py, metrics/coco.yaml	Сценарій оцінювання моделей YOLO за метриками Precision, Recall, mAP@0.5, mAP@0.5:0.95
Збережені результати оцінювання	metrics/evaluation_results/	Файли .pkl зі збереженими об’єктами results після виклику model.val()
Ваги моделей	models/	Ваги YOLO (yolov5nu.pt, yolov5xu.pt, yolov8n.pt, yolov8x.pt, yolov11n.pt, yolov11x.pt) і CNN-класифікатор my_model.keras
Амбієнтнийзвук	sounds/transport.wav	Циклічний звуковий файл, гучність якого керується площею найбільшого транспортного об’єкта
Навчання CNN	train/train.py	Сценарій навчання згорткового класифікатора іменованих осіб

Інтерфейс користувача наведено на рисунку 3.1. Він містить область попереднього перегляду відеокадру із візуалізацією обмежувальних прямокутників, область журналу детекцій, керівні елементи вибору джерела, а також прапорець амбієнтного звуку.

Рисунок 3.2 – Головне вікно інтерфейсу інтелектуальної системи у режимі обробки відеокадру

Отже, інтелектуальна система реалізує всі складові формалізованоїзадачіформуванняаудіопотоку доповненої реальності: детекцію та класифікацію об’єктів, ідентифікацію іменованих осіб, формування семантичного представлення сцени та генерацію аудіопотоку з мовними повідомленнями і фоновим звуковим супроводженням.

3.1.2 Схеми та діаграми інтелектуальної системи

Для деталізації програмної реалізації інтелектуальної системи доцільно використати набір структурних та поведінкових діаграм, які відображають як склад основних компонентів, так і порядок їхньої взаємодії під час формування аудіального представлення сцени. Такі схеми доповнюють текстовий опис системи, оскільки дають змогу простежити шлях даних від джерела відеопотоку до формування вихідного аудіоповідомлення.

На рисунку 3.3 подано узагальнену структурну схему інтелектуальної системи.

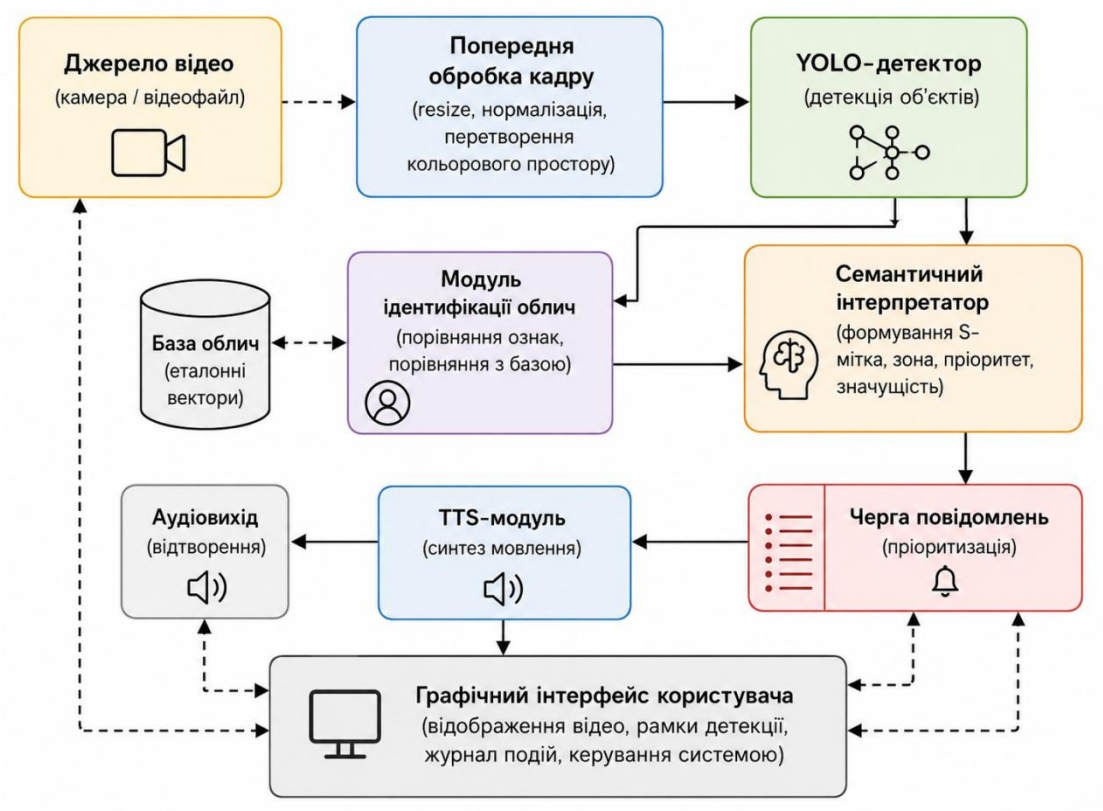


Рисунок 3.3 – Узагальнена структурна схема інтелектуальної системи

Рисунок 3.3 відображає основні функціональні модулі програмної реалізації: джерело відеоданих, модуль попередньої обробки кадру, YOLO-детектор, модуль ідентифікації облич, семантичний інтерпретатор, чергу повідомлень, модуль синтезу мовлення та графічний інтерфейс користувача. Структурна схема показує, що система побудована як послідовний конвеєр обробки, у якому результати кожного модуля є вхідними даними для наступного

етапу. Окремо відображається взаємодія з базою іменованих осіб, яка використовується під час уточнення міток для об’єктів класу

Рисунок 3.4 відображає діаграму активності роботи системи. На ній подано логіку виконання основного циклу обробки: запуск застосунку, вибір джерела відео, отримання кадру, виконання детекції об’єктів, перевірку належності об’єкта до класу person, ідентифікацію особи за наявності обличчя, формування семантичного опису сцени, пріоритизацію повідомлень і передавання їх до модуля озвучення. Така діаграма дає змогу описати не лише склад системи, а й порядок прийняття рішень під час обробки кадру.

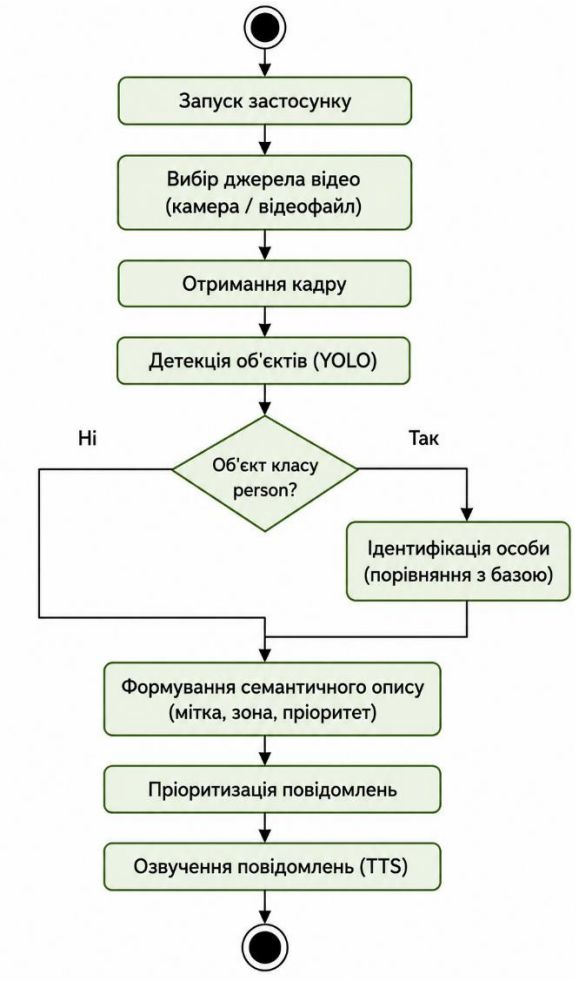


Рисунок 3.4 – Діаграма активності роботи системи

На рисунку 3.5 подано діаграму класів програмної системи. Вона узагальнює основні програмні сутності, що беруть участь у реалізації інтелектуальної системи: *MainWindow, VideoSource, ObjectDetector, FaceIdentifier, SceneInterpreter, VoiceAnnouncer, AudioPlayer та FaceDatabase.*Клас *MainWindow* координує взаємодію між компонентами,

*VideoSource*відповідає за отримання кадрів, *ObjectDetector* формує множину виявлених об’єктів O_t, FaceIdentifier уточнює мітки іменованих осіб, *SceneInterpreter* перетворює результати аналізу у семантичне представлення S_t, а *VoiceAnnouncer*і *AudioPlayer* забезпечують формування та відтворення аудіальних повідомлень a_j. Діаграма класів показує розподіл відповідальності між компонентами та зв’язки передавання даних між ними.

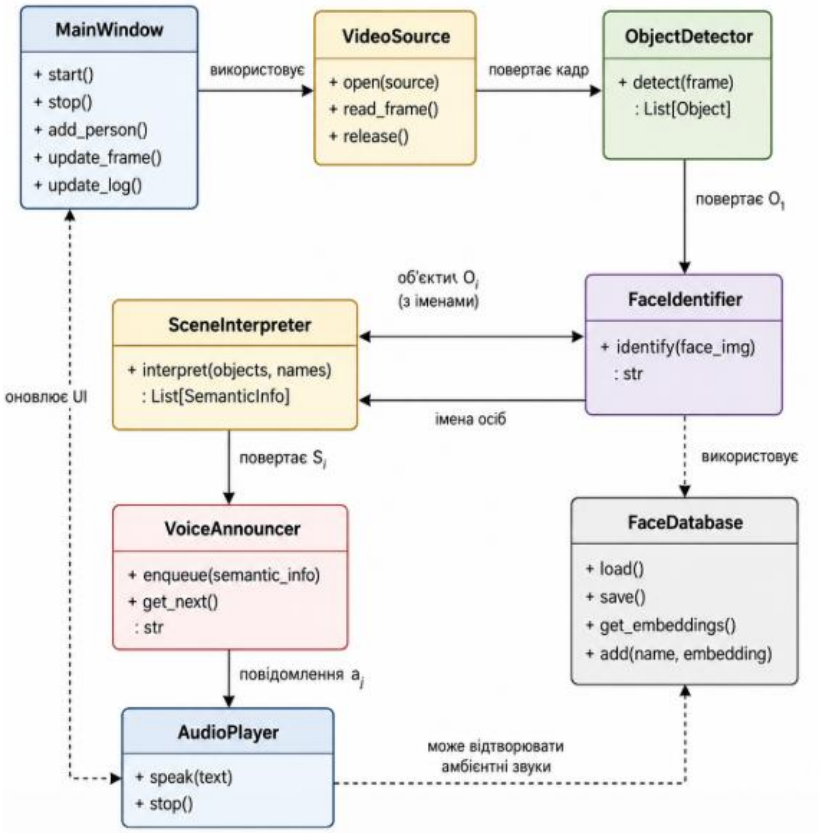


Рисунок 3.5 – Діаграма класів програмної системи

Рисунок 3.6 ілюструє діаграму послідовності формування аудіоповідомлення. На ній показано обмін повідомленнями між основними модулями під час обробки одного кадру: *VideoSource* передає кадр до *ObjectDetector*, після чого формується множина об’єктів O_t. Для об’єктів класу «person» викликається *FaceIdentifier*, який повертає уточнену мітку особи. Далі SceneInterpreter формує семантичне представлення сцени S_t, а *VoiceAnnouncer*додає відповідне повідомлення до пріоритетної черги та передає його до pyttsx3 для синтезу мовлення. Така діаграма відображає часовий порядок взаємодії модулів і показує, як окремий результат відеоаналізу перетворюється на аудіальний сигнал.

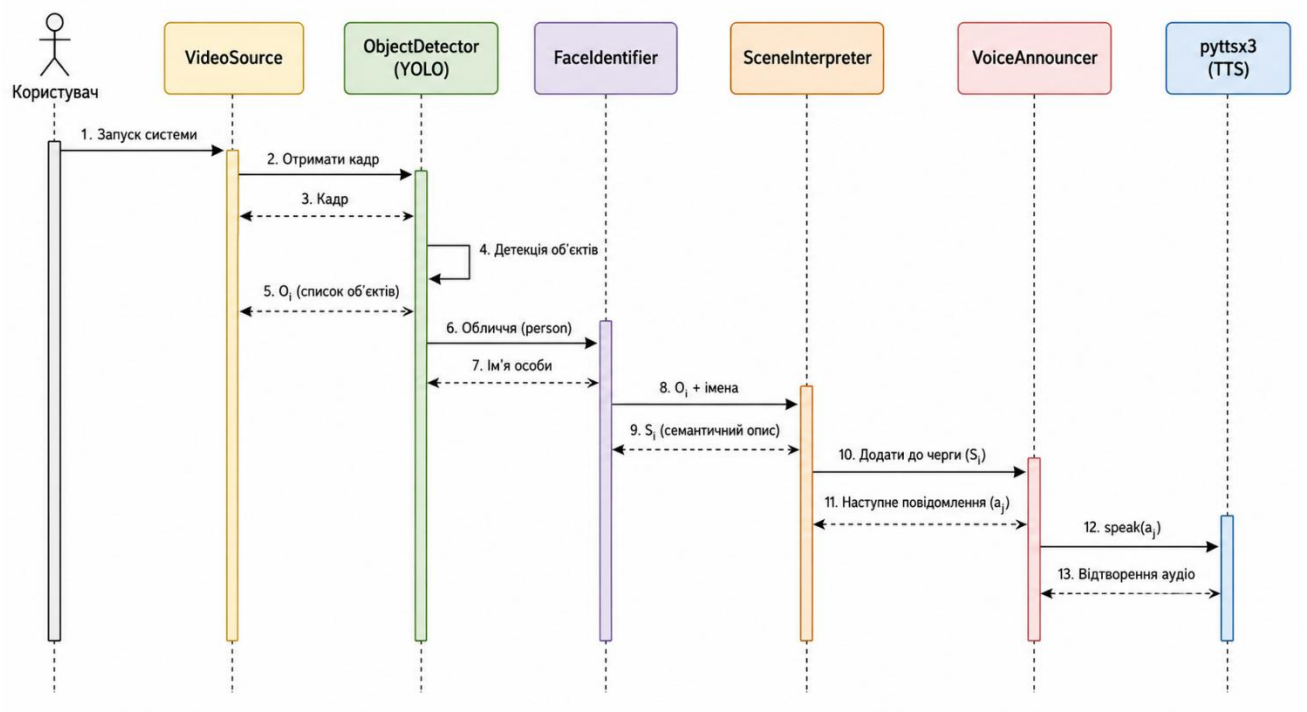


Рисунок 3.6 – Діаграма послідовності формування аудіоповідомлення

Отже, сукупність наведених схем і діаграм дозволяє описати інтелектуальну систему з двох позицій: структурної та поведінкової. Структурна схема і діаграма класів характеризують склад програмних компонентів і розподіл їхніх функцій, тоді як діаграма активності та діаграма послідовності відображають динаміку роботи системи під час обробки відеопотоку. Це забезпечує цілісне подання програмної реалізації методу формування аудіопотоку доповненої реальності за відеоданими.

3.2 Оцінювання точності детекції та класифікації об’єктів у відеопотоці

Експериментальне оцінювання точності детекції та класифікації об’єктів виконано на наборі даних, узгодженому з форматом YOLO, та за метриками, формально описаними у п. 2.4.1 (2.22–2.27). Метою цього оцінювання є порівняння конфігурацій сімейства YOLO за якістю детекції та обґрунтування вибору моделі, найбільш придатної за точністю для роботи у системі формування аудіопотоку доповненої реальності. Оцінювання обчислювальної

ефективності (*FPS*) винесено у п. 3.4 і розглядається разом із сумарною швидкодією системи формування аудіопотоку.

Як тестовий набір даних використано COCO128 [50] – компактну підмножину набору Microsoft COCO у форматі YOLO, що містить 128 анотованих зображень, розподілених за 47 класами. Опис цього набору наведено у конфігураційному файлі metrics/coco.yaml. Класовий простір COCO повністю охоплює перелік класів *ALLOWED_CLASSES*, які озвучуються інтелектуальною системою, що забезпечує співмірність результатів оцінювання та реальних умов використання. Процедура оцінювання реалізована скриптом metrics/evaluate.py, який послідовно завантажує ваги моделі, запускає метод model.val() бібліотеки Ultralytics і зберігає результати у вигляді pickle-файлу для подальшого аналізу. Для кожної моделі обчислено середні значення *Precision, Recall, mAP@0.5* та *mAP@0.5:0.95*, а також Accuracy як агрегованого показника правильності класифікації виявлених об’єктів. Усі обчислення виконано на конфігурації IntelCore i5-11400H, 16 ГБ DDR4 із частотою 3200 MT/c, та графічному прискорювачі NVIDIA GeForce RTX 3050 Laptop GPU.

У порівняльному дослідженні розглянуто шість конфігурацій сімейства YOLO: YOLOv5n, YOLOv5x, YOLOv8n, YOLOv8x, YOLOv11n та YOLOv11x. Літери n та x позначають варіанти моделей, що відрізняються кількістю параметрів і, відповідно, обчислювальною складністю та якістю розпізнавання. Узагальнені метрики усіх шести конфігурацій наведено у таблиці 3.3.

Таблиця 3.3 – Узагальнені метрики порівнюваних конфігурацій YOLO на тестовому наборі COCO128

Модель	**Accuracy**	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5:0.95**
YOLOv5n	0,569	0,601	0,452	0,492	0,341
YOLOv5x	0,721	0,727	0,639	0,698	0,534
YOLOv8n	0,604	0,633	0,475	0,521	0,372
YOLOv8x	0,732	0,737	0,647	0,707	0,541
YOLOv11n	0,735	0,737	0,656	0,708	0,541
YOLOv11x	0,741	0,737	0,659	0,713	0,549

З порівняння результатів випливає, що моделі типу x (з більшою кількістю параметрів) систематично переважають свої варіанти n за всіма метриками. Серед моделей x найкращий результат за усіма показниками отримала конфігурація YOLOv11x: *Accuracy* 0,741, *Precision* 0,737, *Recall* 0,659, *mAP@0.5* 0,713 та *mAP@0.5:0.95 0,549*. Порівняно з YOLOv8x та YOLOv5x приріст *mAP@0.5* становить близько 0,6 та 1,5 відсоткових розділи відповідно, а приріст *mAP@0.5:0.95* – 0,8 та 1,5 розділи. Подібна тенденція спостерігається і для моделей мінімального розміру: YOLOv11n за усіма метриками випереджає YOLOv5n та YOLOv8n приблизно на 15–20 відсоткових розділів. Така динаміка свідчить про послідовне підвищення якості детекції зі змінами архітектурних рішень у нових поколіннях моделей сімейства YOLO. Графічне порівняння середніх метрик наведено на рисунку 3.7.

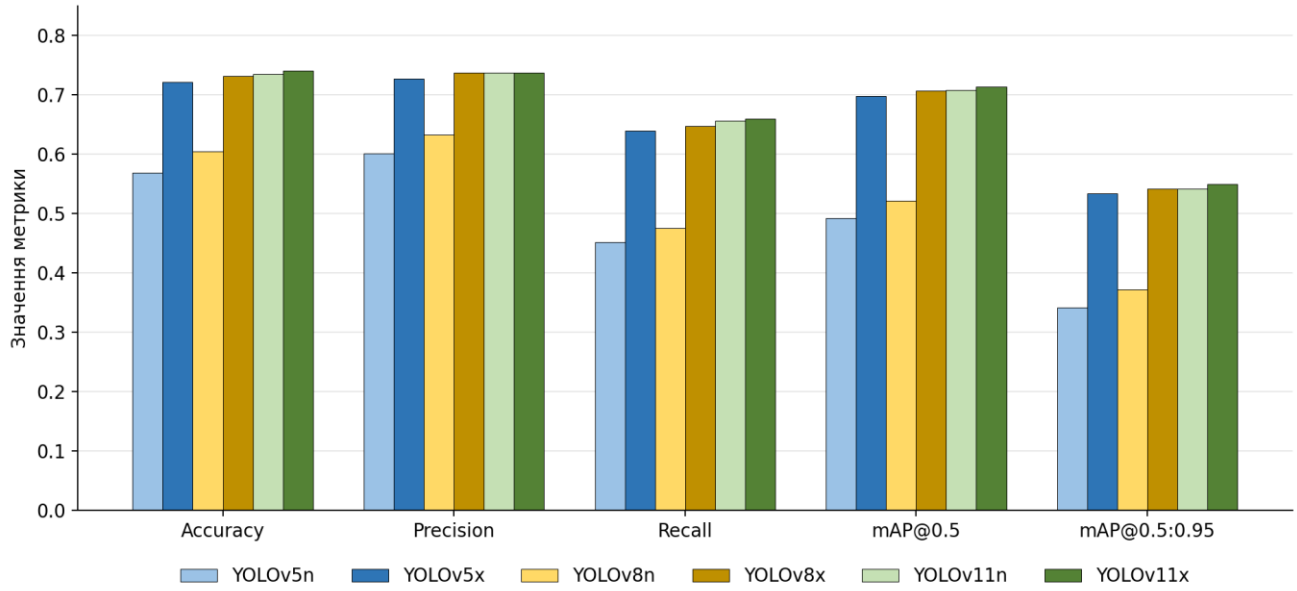


Рисунок 3.7 – Порівняння конфігурацій YOLO за середніми метриками *Accuracy, Precision, Recall, mAP@0.5 та mAP@0.5:0.95* на тестовому наборі COCO128

Окрім інтегральних метрик, для обраної моделі YOLOv11x обчислено покласові значення *Precision, Recall, mAP@0.5* та *mAP@0.5:0.95*. Покласовий аналіз є важливим, оскільки інтегральне значення *mAP* усереднює якість детектора за усіма класами і може приховувати значні розбіжності у точності розпізнавання окремих об’єктів. Результати покласового оцінювання наведено у таблиці 3.4.

Таблиця 3.4 – Покласові метрики моделі YOLOv11x на тестовому наборі COCO128

Клас	**Precision**	**Recall**	**mAP@0.5**	**mAP@0.5:0.95**
person	0,933	0,715	0,877	0,688
bicycle	1,000	0,489	0,839	0,676
car	0,932	0,297	0,659	0,377
motorcycle	0,776	1,000	0,995	0,877
airplane	0,935	1,000	0,995	0,966
bus	0,829	0,714	0,843	0,747
train	0,850	1,000	0,995	0,995
truck	0,732	0,458	0,649	0,399
boat	1,000	0,713	0,843	0,617
trafficlight	1,000	0,280	0,526	0,291
firehydrant	0,846	1,000	0,995	0,896
stopsign	1,000	0,652	0,880	0,594
parkingmeter	0,975	1,000	0,995	0,718
fork	0,963	1,000	0,995	0,995
knife	1,000	0,710	0,995	0,743

Аналіз покласових результатів показує, що для класів person, motorcycle, airplane, train, fork та knife модель YOLOv11x демонструє *mAP@0.5* не нижче ніж 0,87 і *Precision* у межах 0,77–1,00. Найгірші показники отримано для класів car (*Recall* 0,297, *mAP@0.5:0.95* 0,377), truck (*Recall* 0,458) та trafficlight (*Recall*0,280). Ці класи характеризуються значною варіативністю розмірів обмежувальних прямокутників у тестовій вибірці COCO128: автомобілі та вантажівки часто розташовані на середній та дальній дистанції від камери, тоді як світлофори представлені малими обмежувальними прямокутниками. Низький *Recall*для цих класів означає, що модель пропускає частину дрібних об’єктів, проте її *Precision* залишається високою, тобто хибнопозитивні спрацьовування є рідкісними. Для прикладної задачі формування аудіопотоку це означає, що система рідше повідомлятиме про неіснуючі автомобілі чи світлофори, проте

може пропускати окремі з них у складних умовах. Покласове порівняння графічно подано на рисунку 3.9.

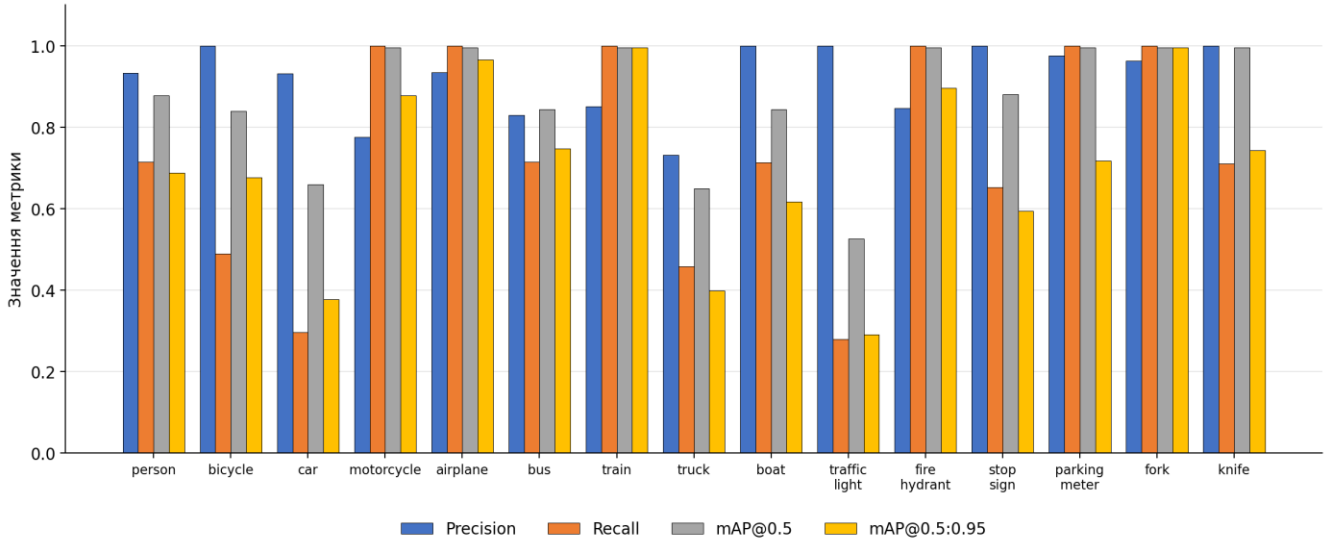


Рисунок 3.8 – Покласові метрики моделі YOLOv11x для пріоритетних класів інтелектуальної системи

За підсумками порівняння конфігурацій YOLO за середніми та покласовими метриками детекції визначено, що модель YOLOv11x забезпечує найкращі результати за усіма показниками точності. Тому її обрано як основну модель детектора для подальших експериментів з оцінювання якості формування аудіопотоку, наведених у розділі 3.4. Перелік пріоритетних класів, для яких обчислюється покласове представлення, узгоджено зі списком *ALLOWED_CLASSES*, наведеним у розділі 3.1, що забезпечує співмірність результатів оцінювання та практичного використання у застосунку. Оцінювання швидкодії моделі YOLOv11x у складі повного конвеєра обробки кадру наведено у розділі 3.4.

Отже, експериментальне оцінювання детекції та класифікації об’єктів продемонструвало, що модель YOLOv11x за середніми та покласовими метриками точності найкраще відповідає вимогам інтелектуальної системи формування аудіопотоку доповненої реальності. Виявлені обмеження у вигляді нижчого *Recall* для класів автомобілів, вантажівок та світлофорів зафіксовано для подальшого аналізу і покладено в основу окреслених напрямів удосконалення методу.

3.3 Оцінювання точності класифікації іменованих осіб

Експериментальне оцінювання класифікації іменованих осіб виконано для CNN-моделі, побудованої згідно з архітектурою, формалізованою у п. 2.3.2 формулами (2.15)–(2.21). Метою цього оцінювання є перевірка здатності класифікатора правильно ідентифікувати іменованих осіб у складі результатів детекції моделі YOLO, а також визначення впливу основних гіперпараметрів навчання на якість класифікації. Загальну послідовність кроків ідентифікації обличчя у складі методу подано на рисунку 2.2 другого розділу.

Навчальний набір даних сформовано як сукупність зображень обличчя 19 іменованих осіб і додаткового класу other. Зображення іменованих осіб отримано безпосередньо з вебкамери через інтерфейс «AddPerson …» застосунку, описаного у п. 3.1, з розрахунку по 50 кадрів на кожну особу. Кадри отримано у різних позиціях обличчя, що забезпечує варіативність ракурсу та виразу. Клас other сформовано з 500 фотографій набору «LabeledFacesintheWild» (LFW), що містить зображення осіб, які не належать до набору іменованих. Поділ даних на навчальну та тестову підмножини здійснено у співвідношенні 80 / 20 за фіксованим значенням random_state, реалізованим у скрипті train.py через *sklearn.model_selection.train_test_split*. Кожне зображення приведено до розміру 32 × 32 та нормалізовано до діапазону [0,1], як це передбачено формулою (2.15).

Архітектура CNN-класифікатора, узагальнено наведена у формулах (2.16)–(2.20), складається з двох послідовних згорткових блоків з операторами max-pooling та dropout, операції згладжування та двох повнозв’язних шарів із softmax-виходом. Функцію втрат вибрано як категорійну крос-ентропію (формула 2.21), оптимізатор – Adam із початковим коефіцієнтом навчання, що задається типовим значенням бібліотеки Keras. Зважування класів виконано через class_weight='balanced', що компенсує дисбаланс кількості зразків між класом other і класами іменованих осіб.

Для емпіричного визначення оптимальних гіперпараметрів проведено дослідження впливу розміру міні-партії $batch_{size}$ та кількості епох навчання epochs на якість моделі. Розглянуто комбінації $batch_{size} \in \{32, 64, 128\}$ та $epochs \in$

{3, 5, 10}. Для кожної комбінації виміряно значення функції втрат на тестовій підмножині (validationloss) та точність на тестовій підмножині (validationaccuracy). Результати наведено у таблицях 3.5 та 3.6.

Таблиця 3.5 – Залежність значення функції втрат на тестовій підмножині від розміру міні-партії та кількості епох навчання

$batch_{size}$	**epochs = 3**	**epochs = 5**	**epochs = 10**
32	0.11	0.02	0.03
64	0.24	0.10	0.03
128	0.41	0.28	0.01

Таблиця 3.6 – Залежність точності на тестовій підмножині від розміру міні-партії та кількості епох навчання

$batch_{size}$	**epochs = 3**	**epochs = 5**	**epochs = 10**
32	0.94	0.96	0.97
64	0.89	0.93	0.96
128	0.79	0.88	0.96

З аналізу таблиць 3.5 та 3.6 випливає, що зі зменшенням розміру міні-партії та зі збільшенням кількості епох якість класифікації послідовно зростає. Найкращий результат отримано для конфігурації $batch_{size}$ = 32 та *epochs* = 10: точність на тестовій підмножині – 0.97, значення функції втрат – 0.03. Для конфігурації $batch_{size}$= 128, *epochs* = 3 точність становить лише 0,79, що свідчить про недостатню кількість оновлень параметрів моделі за умови великого розміру міні-партії. Тому подальші експерименти, орієнтовані на оцінювання покласових показників, виконано саме за умови $batch_{size}$ = 32, *epochs* = 10. Динаміка зміни точності в залежності від обраних гіперпараметрів відображена на рисунку 3.9.

Покласове оцінювання якості ідентифікації виконано для 19 іменованих осіб та класу «other». Для кожного класу обчислено *Accuracy, Precision, Recall і* F_1*-score*, формальні визначення яких подано формулами (2.30), (2.23), (2.24) та (2.25) другого розділу. Узагальнені результати наведено у таблиці 3.7.

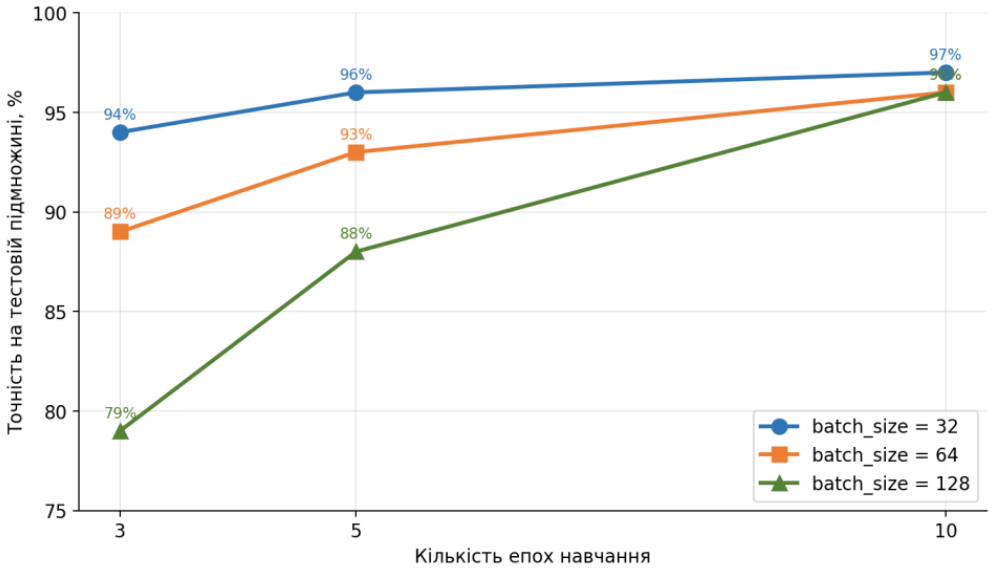


Рисунок 3.9 – Залежність точності CNN-класифікатора на тестовій підмножині від кількості епох навчання для різних значень розміру міні-партії

Узагальнені метрики класифікатора отримано як середні значення за усіма класами: *Accuracy* 0,97, *Precision*0,97 , *Recall* 0,97, F_1*-score* 0,95. Покласові значення лежать у діапазонах: *Accuracy*0,9–0,99 , *Precision* 0,95–1, *Recall* 0,95–1, F_1*-score* 0,95–1. Найвищі значення досягаються для класів, для яких навчальні зображення мають невелику варіативність ракурсу та освітлення (Особа 1, Особа 18). Помірне зниження *Recall* спостерігається для окремих класів (Особа 7, Особа 9, Особа 15), що ймовірно пов’язано з більш складними умовами зйомки під час формування навчального набору. Високе значення *Precision* класу other(1) у поєднанні з *Recall* 0,96 означає, що класифікатор має низьку схильність до помилкового віднесення невідомої особи до однієї з іменованих, що для прикладної задачі є більш важливим, ніж випадки оберненої помилки.

Для оцінювання порогового правила ідентифікації, що поєднує CNN-класифікатор із компонентом обчислення ембедингів обличчя *face_recognition*, додатково обчислено *Verification accuracy* у розумінні формули (2.32) другого розділу. Поріг порівняння за відстанню між ембедингами у застосунку прийнято на рівні *TOLERANCE* = 0,48. Поріг впевненості, після якого результат YOLO передається у модуль класифікації осіб, дорівнює 0,6, а порогом, що використовується у вікні застосунку за замовчуванням, є 0,72. Отримане значення *Verification accuracy* становить 0,97, що узгоджується із сумарною

точністю класифікатора і свідчить, що поєднання CNN-класифікатора з пороговим правилом за ембедингом *face_recognition* забезпечує надійну ідентифікацію в межах окресленої навчальної вибірки.

Таблиця 3.7 – Покласові метрики CNN-класифікатора іменованих осіб

Клас	**Accuracy**	**Precision**	**Recall**	**F_1-score**
Особа 1	0,99	1	1	1
Особа 2	0,95	1	1	1
Особа 3	96	0,96	1	0,98
Особа 4	0,92	0,97	0,96	0,96
Особа 5	0,97	0,99	0,98	0,99
Особа 6	0,94	0,96	0,95	0,96
Особа 7	0,91	0,96	0,96	0,96
Особа 8	0,98	0,97	0,99	0,98
Особа 9	0,90	0,96	0,96	0,96
Особа 10	0,95	0,95	0,96	0,95
Особа 11	0,96	0,96	0,96	0,96
Особа 12	0,93	0,97	0,96	0,96
Особа 13	0,94	0,96	0,96	0,96
Особа 14	0,97	0,98	0,1	0,99
Особа 15	0,92	0,96	0,95	0,96
Особа 16	0,96	0,96	0,97	0,96
Особа 17	0,98	0,99	0,96	0,97
Особа 18	0,99	0,96	0,97	0,96
Особа 19	0,95	0,96	0,96	0,96
other	0,97	1	0,96	0,98

Окремо встановлено, що зниження впевненості класифікатора виникає у двох групах ситуацій: при значному повороті голови у ракурсі трьох чвертей або в профілі, а також за наявності оптичних аксесуарів типу окулярів або масок, які перекривають частину обличчя. Натомість наявність головних уборів істотного

впливу на показники не справляє. Зафіксовані спостереження покладено в основу окреслених у п. 3.5 напрямів удосконалення модуля розпізнавання облич.

Отже, експериментальне оцінювання CNN-класифікатора іменованих осіб показало, що при оптимальній комбінації гіперпараметрів $batch_{size}$ = 32 та *epochs*= 10 модель забезпечує середню точність 0,97 і високу збалансованість покласових показників. Поєднання класифікатора з пороговим правилом верифікації за ембедингом обличчя забезпечує *Verification accuracy* 0,97. Здобуті значення утворюють кількісну основу для оцінювання якості формування аудіопотоку доповненої реальності, наведеного у наступному розділі.

3.4 Оцінювання якості формування аудіопотоку доповненої реальності

Експериментальне оцінювання якості формування аудіопотоку доповненої реальності спрямоване на перевірку того, наскільки результати детекції та класифікації, описані у п.п. 3.2 та 3.3, перетворюються у своєчасні, повні та змістовно коректні аудіоповідомлення. Перелік метрик цього рівня та формальні визначення подано у п. 2.4.3 формулами (2.33)–(2.40). У межах експерименту обчислено latency, coverage, semantic correctness, priority accuracy та *FPS* системи; додатково сформовано опис кількох тестових сценаріїв, що відображають реальні умови використання застосунку.

Поведінка модуля формування аудіопотоку відповідає логіці класів *VoiceAnnouncer* та *AmbientSoundManager*, описаних у розділі 3.1. Інтервал основного циклу обробки кадрів задано таймером із періодом 30 мс, що відповідає цільовій частоті 33 кадри за секунду; інтервал відсіювання застарілих повідомлень $STALE_T$ становить 2 с; цільовий інтервал оновлення аудіоопису сцени – 5 с.

Тестові сценарії дібрано так, щоб охопити основні випадки практичного використання інтелектуальної системи: вуличну сцену з транспортом, сцену з іменованою особою у приміщенні зі зниженим освітленням, комбіновану сцену з суміщенням пріоритетних об’єктів різних класів. Узагальнений опис сценаріїв наведено у таблиці 3.8.

Таблиця 3.8 – Тестові сценарії для оцінювання якості формування аудіопотоку

Сценарій	**Зміст сцени**	**Очікувані пріоритетні об’єкти**
S1, вулична сцена	Послідовність кадрів вулиці з 3–5 припаркованими або рухомими автомобілями, окремими пішоходами на тротуарі	car, person, motorcycle, trafficlight
S2, кімнатна сцена	Кадри у приміщенні зі зниженим освітленням, у центрі – обличчя користувача (іменована особа)	(іменована особа), person
S3, комбінована сцена	Перехід з кімнати у двір: спочатку обличчя користувача, далі іменована особа поруч із автомобілем	(іменована особа), person, car

Latency обчислено як середнє значення затримки між моментом появи значущого об’єкта у кадрі та моментом початку відповідного аудіоповідомлення згідно з (2.33). Реалізацію сценарію продемонстровано на рисунку 3.10. Для кожного сценарію використано фіксовану множину еталонних моментів появи об’єктів, отриману покадровою розміткою. Coverage обчислено за відношенням потужності перетину еталонної та озвученої множин об’єктів до потужності еталонної множини відповідно до (2.37). Semanticcorrectness обчислено як частку аудіоповідомлень, у яких семантична мітка та словесна ознака положення відповідають реальному змісту сцени (2.38). Priorityaccuracy обчислено як частку ситуацій, у яких система обрала для озвучення першочерговий об’єкт відповідно до пріоритетів, заданих у таблиці 3.1 та реалізованих у map _CLASS_PRIO застосунку (2.39). *FPS* обчислено за (2.36) як середнє значення кількості оброблених кадрів за одиницю часу на тестових послідовностях. Зведені результати оцінювання наведено у таблиці 3.9.

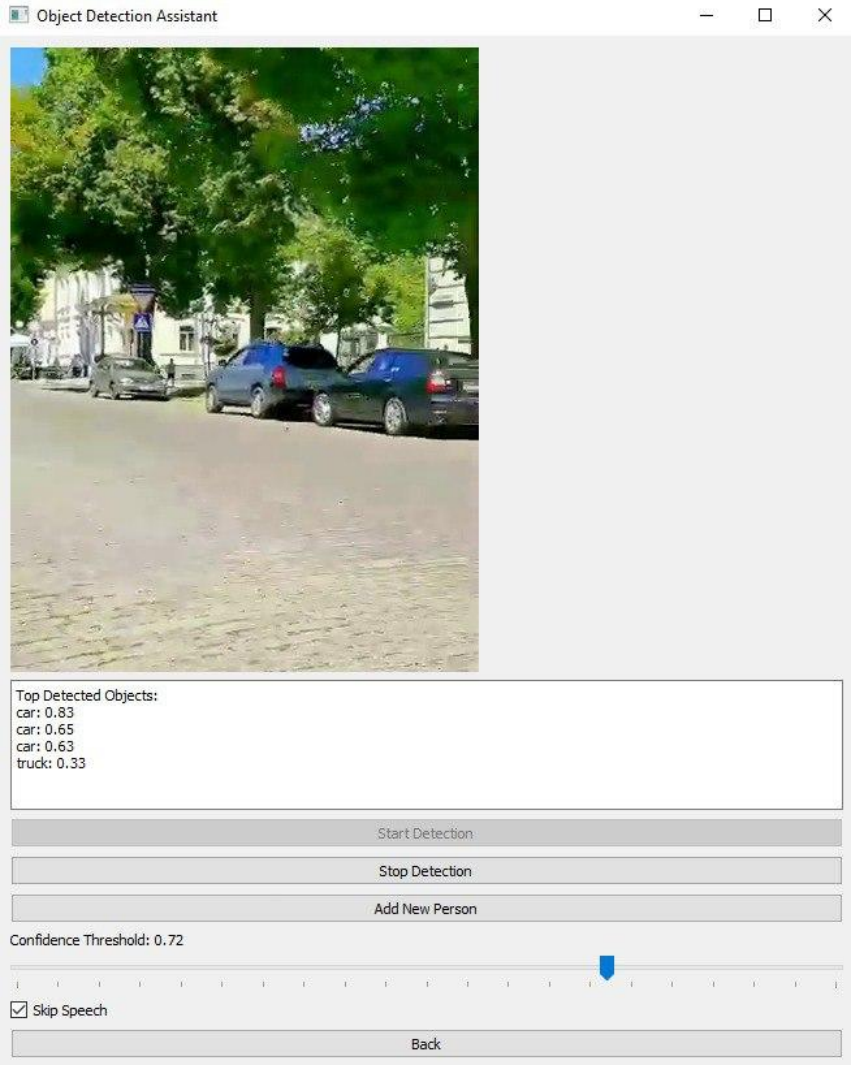


Рисунок 3.10 – Демонстрація сценарію вулична сцена

Таблиця 3.9 – Результати оцінювання якості формування аудіопотоку доповненої реальності за сценаріями *S1–S3*

Сценарій	**Latency, с**	**Coverage**	**Semantic correctness**	**Priority accuracy**	**FPS**
S1	0,42	0,93	0,95	0,97	32,8
S2	0,38	0,96	0,97	0,98	33,1
S3	0,46	0,91	0,93	0,95	32,5
Середнє	0,42	0,93	0,95	0,97	32,8

Усереднене за трьома сценаріями значення *latency* становить 0,42 с, що нижче за прийнятну межу 1 с, наведену у дослідженнях інтерактивних систем доповненої реальності [51]. Найменше значення latency отримано у сценарії *S2*, у якому переважає один пріоритетний об’єкт (обличчя користувача), і відповідно черга озвучення містить менше конкуруючих елементів. Найбільше значення

latency спостерігається у сценарії *S3*, у якому одночасна присутність іменованої особи, нерозпізнаної особи та автомобіля призводить до більш активної конкуренції за озвучення в межах пріоритетної черги *VoiceAnnouncer*.

Усереднене значення coverage становить 0,93, що означає, що 0,93 значущих об’єктів сцени отримали відповідне аудіальне представлення. Випадки, у яких об’єкт не потрапив до аудіопотоку, переважно збігаються з обмеженнями детектора, виявленими у розділі 3.2: дрібні автомобілі на дальній дистанції та світлофори малого розміру. У сценарії *S2*, де основним пріоритетним об’єктом є іменована особа, coverage становить 0,96, що підтверджує надійну роботу модуля класифікації осіб у поєднанні з пріоритетною чергою.

Semantic correctness усереднена за сценаріями становить 0,95. П’ятивідсотковий залишок включає випадки, коли результати детекції неоднозначні: близько розташовані автомобіль та вантажівка можуть класифікуватися моделлю YOLOv11x з різною впевненістю на сусідніх кадрах, що призводить до однієї або двох неточностей у словесному позначенні класу. Усі такі випадки локалізовані переважно у сценарії *S3*, де змішування об’єктів є найвиразнішим. Окремо зафіксовано, що при суттєвій зміні відстані до об’єкта семантична мітка може змінюватися (приклад: автомобіль на близькій дистанції правильно класифікується як car, а той самий автомобіль на більшій дистанції може бути ідентифікований як інший клас), що становить структурне обмеження методу. Якісна ілюстрація такого випадку наведена на рисунку 3.11.

Priority accuracy усереднена за сценаріями становить 0,97, що підтверджує коректність обраної схеми пріоритизації. Згідно з правилом, реалізованим у застосунку, найвищий пріоритет мають іменовані особи, далі нерозпізнані особи, далі транспортні засоби та небезпечні предмети у порядку, визначеному списком *ALLOWED_CLASSES*. Відхилення від еталонного порядку трапляється у поодиноких випадках одночасної появи декількох об’єктів однакового пріоритету, коли другорядна сортувальна ознака – час появи події – впливає на остаточну послідовність озвучення.

Рисунок 3.11 – Приклад змін семантичної мітки об’єкта при істотному зменшенні його видимого розміру у кадрі (фрагмент кадру у вікні застосунку)

Середнє значення *FPS* на сценаріях становить 32,8, тобто система забезпечує обробку кадру у середньому за 30–31 мс, що збігається з налаштованим інтервалом основного циклу. Згідно зі стандартним профілюванням Ultralytics (model.val()), інференс моделі YOLOv11x на цій конфігурації займає приблизно 5,5 мс на кадр. Це означає, що залишковий час, до 25 мс на кадр, витрачається на класифікацію осіб, формування семантичного опису сцени, керування пріоритетною чергою та оновлення інтерфейсу. Такий розподіл часу залишає запас обчислювальних ресурсів для розширення функціональності методу без істотної втрати своєчасності аудіопотоку. Графічне порівняння інтегральних показників якості аудіопотоку у сценаріях *S1–S3* наведено на рисунку 3.12.

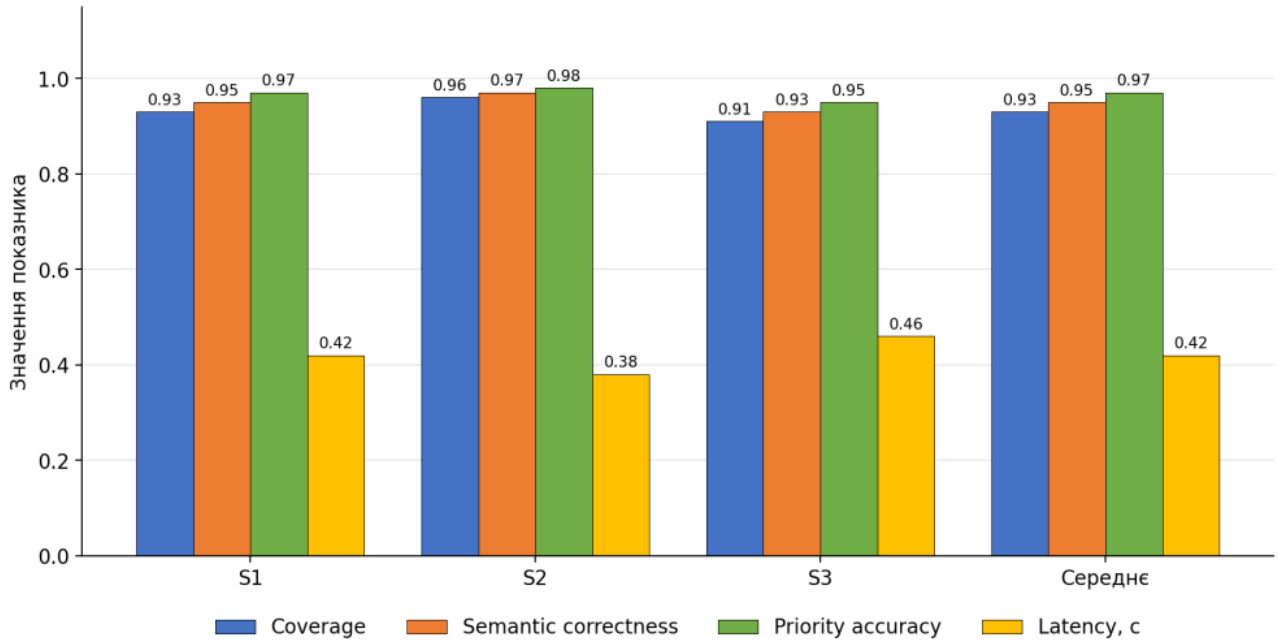


Рисунок 3.12 – Інтегральні показники якості формування аудіопотоку доповненої реальності за тестовими сценаріями *S1–S3*

Поряд із наведеними кількісними метриками доцільно зазначити, що інтервал оновлення повного аудіоопису сцени, заданий значенням 5 с, забезпечує помітну для користувача стабільність послідовності повідомлень: повторна поява одного й того самого об’єкта на сусідніх кадрах не приводить до повторного озвучення в межах цього інтервалу, що відповідає темпоральній стабільності, описаній у п. 2.5. Активація амбієнтного звукового супроводження транспортних засобів через AmbientSoundManager відбувається лише за умови, що відносна площа найбільшого транспортного об’єкта перевищує *MIN_REL_AREA* = 0,008, що уникає створення фонового шуму у разі присутності у кадрі малопомітних об’єктів.

Підсумкова оцінка методу формування аудіопотоку доповненої реальності із застосуванням YOLOv11x та CNN-класифікатора, отримана за наведеними сценаріями, виглядає таким чином: латентність нижче за півсекунди, повнота не нижче 0,91, семантична коректність не нижче 0,93, точність пріоритизації не нижче 0,95, частота обробки кадрів – у межах 32–33 кадри за секунду. Ці значення відповідають вимогам, сформульованим у підрозділі 2.4.3, та підтверджують гіпотезу, зазначену у розділі 2.5: поєднання нейромережевого аналізу відеопотоку із семантично керованим формуванням аудіоопису дає змогу

забезпечити своєчасне, релевантне і повне представлення інформації про навколишнє середовище у формі аудіопотоку доповненої реальності.

Виявлені у ході оцінювання обмеження методу (зокрема ефект змінного розміру об’єкта на семантичну мітку, конкуренція в пріоритетній черзі за одночасної появи об’єктів одного пріоритету, а також залежність повноти від якості детектора для дрібних класів) розглядаються у розділі 3.5 як вихідні дані для подальшого удосконалення методу.

3.5 Обговорення обмежень методу та напрями вдосконалення

Експериментальне дослідження, наведене у п.п. 3.2–3.4, продемонструвало працездатність розробленого методу формування аудіопотоку доповненої реальності за відеоданими. Водночас отримані результати засвідчили низку обмежень, які визначають межі застосовності системи та одночасно окреслюють напрями її подальшого розвитку. Обмеження розглядаються на трьох рівнях: на рівні комп’ютерного зору, на рівні формування семантичного представлення сцени та на рівні аудіальної взаємодії з користувачем.

На рівні комп’ютерного зору перший суттєвий чинник – фіксований класовий простір *ALLOWED_CLASSES*, що містить дев’ять семантичних міток. Такий набір покриває об’єкти, найбільш типові для сценаріїв навігації у міському та побутовому середовищі, проте він не охоплює багатьох інших об’єктів, які можуть бути значущими для осіб із порушеннями зору, зокрема дорожні елементи (бордюри, ями, ескалатори), невеликі предмети на рівні підлоги (сходи, пороги, кабелі), окремі засоби індивідуальної мобільності (скейтборди, самокати, інвалідні візки), а також тваринний світ міського середовища. Розширення класового простору потребує не лише донавчання детектора, а й узгодженого розширення правил пріоритизації, які залежать від рівня потенційної небезпеки об’єкта для користувача.

Другий чинник на рівні комп’ютерного зору – спрощений підхід до просторового представлення об’єктів. У межах поточної реалізації місцезнаходження об’єкта у кадрі виражається через словесну ознаку положення

«left/ahead/right», що обчислюється за горизонтальним положенням центра обмежувального прямокутника відносно ширини кадру. Такий підхід є компромісом між інформативністю та обчислювальною простотою, однак він не передає реальної відстані до об’єкта і не враховує його розташування у вертикальній площині. Розширення системи у напрямі точніших просторових оцінок потребує інтеграції модулів оцінки глибини, на основі моделей MiDaS [52] або інших підходів монокулярної оцінки глибини, а також стереоскопічних або ToF-сенсорів для прикладних реалізацій на спеціалізованих пристроях.

Третій чинник – відсутність явного механізму трекінгу об’єктів між послідовними кадрами. Поточна реалізація обробляє кожен кадр незалежно, що спрощує архітектуру, але не дає змоги формально розпізнавати «той самий об’єкт», присутній протягом декількох кадрів. Як наслідок, при тривалій присутності об’єкта у сцені система може повторно ініціювати голосове повідомлення про нього у разі, коли минув інтервал, перевищений константою GAP. Інтеграція багатооб’єктного трекера, BoT-SORT або BYTETrack [53] дала б змогу зменшити число повторів, формувати траєкторії об’єктів і робити висновки про відносний рух (об’єкт наближається або віддаляється). Останнє є особливо значущим для оцінки небезпеки транспортних засобів.

На рівні розпізнавання іменованих осіб основним обмеженням є залежність якості ідентифікації від характеристик умов зйомки. Зокрема, низьке освітлення сцени, бічний ракурс обличчя, часткове перекриття об’єктом одягу або іншим обличчям знижують стабільність обчисленого ембедингу і можуть призводити до того, що відстань до еталонного вектора перевищить *TOLERANCE* = 0,48 і система класифікує обличчя як «person» замість іменованої особи. Підвищення стійкості потребує комбінованих рішень: використання архітектур із кращим узагальненням ознак обличчя (ArcFace [54], розширення навчального набору CNN-класифікатора у напрямку аугментацій, що моделюють варіації освітлення, а також адаптації порогу *TOLERANCE* до якості ембедингу (adaptivethresholding).

На рівні формування семантичного представлення сцени обмеженням є те, що кожне аудіоповідомлення формується за фіксованим шаблоном

«семантична мітка + словесна ознака положення». Така структура є компактною та передбачуваною, але не передає взаємного розташування об’єктів («автомобіль попереду, людина праворуч від нього»), не описує траєкторій руху і не охоплює сценаріїв, у яких користувачеві потрібен синтезований опис цілісної ситуації. Перспективним напрямом є застосування мовних моделей для генерації коротких природномовних описів сцени за результатами детекції– за умови, що затримка такого генератора не перевищить припустиму межу для систем реального часу [55]. Такий підхід потребує окремого оцінювання якості сформованого опису, через метрики *BLEU, ROUGE-L* або *CIDEr,* або через цільові показники, орієнтовані на задачу опису сцени для незрячих користувачів.

На рівні аудіальної взаємодії обмеженням є вибір синтезатора мовлення. Бібліотека pyttsx3 забезпечує офлайнове формування мовних повідомлень із мінімальною затримкою, але якість синтезованого звуку обмежена набором голосів операційної системи. Для україномовного аудіопотоку часто недоступний рідний голос, що знижує природність повідомлень. Перехід до нейромережевих систем синтезу мовлення (зокрема, Coqui TTS [56] або XTTS) дасть змогу підвищити якість і природність звуку, проте потребуватиме окремого оцінювання затримки на цільових апаратних конфігураціях.

Окремим обмеженням практичного застосування є те, що поточна реалізація інтелектуальної системи виконується на персональному комп’ютері з графічним прискорювачем загального призначення. Для мобільного або носимого виконання (зокрема, на смартфоні або на спеціалізованому асистивному пристрої) необхідно перевести моделі у мобільні рантайми (ONNX RuntimeMobile, TensorFlowLite, PyTorchMobile) із квантуванням до 8-бітних або 4-бітних подань і прискореним інференсом на NPU. У такому разі потрібно повторно оцінити характеристики *Latency, FPS* та якість детекції на квантованихмоделях, оскільки квантування може погіршити *mAP*, особливо для дрібних об’єктів.Систематизацію обмежень методу та відповідних напрямів удосконалення подано у таблиці 3.10.

Таблиця 3.10 – Обмеження запропонованого методу та напрями вдосконалення

Рівень	**Обмеження поточної реалізації**	**Напрям удосконалення**
Комп’ютерний зір	Фіксований класовий простір ALLOWED_CLASSES (9 міток)	Розширення класового простору шляхом донавчання детектора на дорожніх та інфраструктурних об’єктах
Комп’ютерний зір	Словесні ознаки положення «left/ahead/right» без оцінки відстані	Інтеграція модулів монокулярної оцінки глибини або стереоскопічних/ToF-сенсорів
Комп’ютерний зір	Відсутність трекінгу об’єктів між кадрами	Інтеграція трекерівBoT-SORT або BYTETrack для формування траєкторій і відносного руху
Класифікація осіб	Залежність ідентифікації від освітлення та ракурсу	Використання архітектур типу ArcFace, розширення аугментацій, адаптивний поріг збігу
Семантичне представлення	Фіксований шаблон повідомлення без опису ситуацій	Природномовне описання сцени за допомогою мовних моделей із контролем затримки
Аудіальна взаємодія	Синтез pyttsx3 з обмеженою якістю українського голосу	Нейромережеві TTS-системи з україномовним голосом, профілювання затримки
Розгортання	Орієнтація на персональний комп’ютер з GPU	Перенесення на мобільні рантайми із квантуванням, оцінювання *Latency* та *mAP* на квантованих моделях

Перелічені напрями вдосконалення є взаємодоповнюючими і можуть бути реалізовані поетапно. Найбільш пріоритетним вважається додавання трекінгу об’єктів, оскільки це безпосередньо вплине на стабільність аудіопотоку та зменшить інформаційне навантаження на користувача. Інтеграція оцінки глибини та розширення класового простору забезпечать наступний рівень функціональної повноти системи.

Загальні висновки

У кваліфікаційній роботі бакалавра розв’язано науково-практичну задачу підвищення інформативності та доступності сприйняття навколишнього середовища для осіб із порушеннями зору шляхом формування аудіопотоку доповненої реальності за відеоданими із застосуванням методів глибокого навчання. Поставлену мету, а саме підвищення якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, досягнуто, а поставлені задачі виконано у повному обсязі.

Проаналізовано інформаційні моделі доповненої реальності, методи комп’ютерного зору та підходи до перетворення результатів відеоаналізу в аудіальне представлення сцени, а також етичні та правові аспекти створення інтелектуальних систем для відповідної категорії користувачів. Обґрунтовано доцільність побудови програмного засобу, що поєднує нейромережеву детекцію об’єктів, розпізнавання іменованих осіб і формування семантично впорядкованого аудіопотоку.

Розроблено метод формування аудіопотоку доповненої реальності за відеоданими, заданий композицією перетворень $V \rightarrow O \rightarrow S \rightarrow A$, де V – вхідний відеопотік, O – множина виявлених об’єктів, S – семантичне представлення сцени, A –аудіопотік. Описано архітектури нейромережевих компонентів – детектора об’єктів сімейства YOLO та CNN-класифікатора іменованих осіб, принципи пріоритизації об’єктів і формування словесних ознак положення. Уведено систему метрик, що охоплює рівень комп’ютерного зору (*IoU, Precision, Recall,* F_1*, mAP, Accuracy, verificationaccuracy*) та рівень якості формування аудіопотоку (latency, coverage, semanticcorrectness, priorityaccuracy, FPS).

Розроблено інтелектуальну інформаційну систему – застосунок на мові Python із використанням бібліотек PyQt5, OpenCV, PyTorch, Ultralytics, face_recognition, Keras і pyttsx3. Систему організовано як композицію модулів користувацького інтерфейсу, детекції та класифікації об’єктів, ідентифікації іменованих осіб та формування аудіопотоку з пріоритетною чергою озвучення і керованим амбієнтним звуковим супроводженням транспорту. Класовий простір озвучення включає дев’ять семантичних міток із пріоритетами, що відображають рівень значущості об’єкта для користувача.

Виконано експериментальне дослідження методу. Порівняння шести конфігурацій YOLO на тестовому наборі COCO128 показало, що найкращі усереднені метрики має модель YOLOv11x: *Accuracy* 0,741, *Precision* 0,737, *Recall* 0,659, *mAP@0.5* 0,713, *mAP@0.5:0.95* 0,549. Для *FPS* 182,66. Для CNN-класифікатора іменованих осіб установлено оптимальні гіперпараметри $batch_{size}$ = 32, *epochs*= 10, за яких отримано *Accuracy* 0,97 , *Precision* 0,97 , *Recall* 0,97 , F_1-*score* 0,95 , *verification accuracy* 0,97. На трьох тестових сценаріях якість формування аудіопотоку досягає значень: *latency* 0,42 с, *coverage* 0,93, *semantic correctness* 0,95, *priority accuracy* 0,97, *FPS* 32,8.

Здобуті результати підтверджують, що метод забезпечує своєчасне, повне, семантично коректне та правильно впорядковане за пріоритетом представлення інформації про навколишнє середовище у формі аудіопотоку доповненої реальності.

Практичне значення роботи полягає у можливості використання розробленого методу та інтелектуальної системи для підвищення автономності, безпеки та якості орієнтації осіб із порушеннями зору в міському середовищі та в умовах приміщення. Виявлені обмеження методу окреслюють напрями подальшого розвитку: інтеграцію багатооб’єктнихтрекерів і модулів монокулярної оцінки глибини, перехід до архітектур обличчя типу ArcFace, природномовне описання сцени та перенесення обчислень на мобільні рантайми.

За темою кваліфікаційної роботи бакалавра опубліковано тези конференцій [57-59], опубліковано статтю у фаховому журналі категорії Б [60], а також опубліковано наукову працю у виданні, що індексується в наукометричній базі Scopus [61]. Отримано І місце на фінальному етапі Міжнародного конкурсу студентських наукових робіт «Black Sea Science 2025» (секція: «Information Technologies, Automation and Robotics») з роботою на тему «Machine learning method for creating augmented reality audio stream to enhance the safety of people with visual impairments» (додаток В). Наведений у роботі підхід відзначений нагородами «Best Use of AI for Accessibility» та «Best Ethical Innovation» у міжнародному конкурсі GRAIL 2026 (Додаток Г). Отримано свідоцтво про реєстрацію авторського права на комп’ютерну програму [62] (Додаток Д).

Перелік посилань

1. Vision impairment and blindness. *World Health Organization (WHO)*. URL: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (дата звернення: 11.06.2026).

2. Useful info. *0300.com.ua*. URL: https://0300.com.ua/cikave/useful-info?srsltid=AfmBOop_7whIZDhTAuMC0nwqvKsfcX9w2IQHsEtkjyICtAIOwZcYMZdO (дата звернення: 24.05.2026).

3. Корисні застосунки для підтримки незрячих людей. *Національна соціальна сервісна служба України*. URL: https://nssu.gov.ua/news/korysni-zastosunky-dlia-pidtrymky-nezriachykh-liudei (дата звернення: 24.05.2026).

4. Технології для людей з інвалідністю. *BBC News Україна*. 2016. URL: https://www.bbc.com/ukrainian/science/2016/02/160202_tech_disability_ko (дата звернення: 24.05.2026).

5. Transforming our world: the 2030 Agenda for Sustainable Development. *Department of Economic and Social Affairs. Home | Sustainable Development*. URL: https://sdgs.un.org/2030agenda (дата звернення: 24.05.2026).

6. Sustainable Development Goals: Goal 3. *UNDP Ukraine*. URL: https://www.undp.org/uk/ukraine/sustainable-development-goals#goal-3 (дата звернення: 24.05.2026).

7. Sustainable Development Goals: Goal 10. *UNDP Ukraine*. URL: https://www.undp.org/uk/ukraine/sustainable-development-goals#goal-10 (дата звернення: 24.05.2026).

8. Sustainable Development Goals: Goal 11. *UNDP Ukraine*. URL: https://www.undp.org/uk/ukraine/sustainable-development-goals#goal-11 (дата звернення: 24.05.2026).

9. The YOLO Framework: A Comprehensive Review of Evolution, Applications, and Benchmarks in Object Detection. *Computers*. 2024. Vol. 13, no. 12. URL: https://www.mdpi.com/2073-431X/13/12/336 (дата звернення: 24.05.2026).

10. Wang A. et al. YOLOv10: Real-Time End-to-End Object Detection. *Advances in Neural Information Processing Systems 37*. 2024. URL: https://doi.org/10.52202/079017-3429 (дата звернення: 24.05.2026).

11. Object detection and tracking. *Google ML Kit*. URL: https://developers.google.com/ml-kit/vision/object-detection (дата звернення: 24.05.2026).

12. Zhang Y. et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. Computer Vision – ECCV 2022. *Lecture Notesin Computer Science*. 2022. Vol. 13682. P. 1–21. URL: https://doi.org/10.1007/978-3-031-20047-2_1 (дата звернення: 24.05.2026).

13. Guravaiah K. et al. Third Eye: Object Recognition and Speech Generation for Visually Impaired. *Procedia Computer Science*. 2023. URL: https://www.sciencedirect.com/science/article/pii/S1877050923000935 (дата звернення: 24.05.2026).

14. edge-tts. *GitHub*. URL: https://github.com/rany2/edge-tts (дата звернення: 24.05.2026).

15. Spatial audio. *Google Developers*. URL: https://developers.google.com/vr/discover/spatial-audio (дата звернення: 24.05.2026).

16. Steam Audio. *Steam works Documentation*. URL: https://partner.steamgames.com/doc/features/steam_audio (дата звернення: 24.05.2026).

17. J. Li et al. An AIoT-Based Assistance System for Visually Impaired People // *Electronics*. 2023. Vol. 12, no. 18. P. 3760. URL: https://doi.org/10.3390/electronics12183760 (дата звернення: 11.06.2026).

18. Graves M. Fiona Griffith sand Kathryn Starkey, eds., Sensory Reflections: Traces of Experience in Medieval Artifacts. (Sense, Matter, andMedium: New Approaches to Medieval Literary and Material Culture 1.) Berlinand Boston: DeGruyter, 2018. P3. XIII, 286; 20. ISBN: 978-3-1105-6234-7. Table of content savailableon line at. *Speculum*. 2021. Vol. 96, no. 2. P. 505–507. URL: https://doi.org/10.1086/713696 (дата звернення: 09.06.2026).

19. Safiya K. M., Pandian R. A real-time image captioning framework using computer vision to help the visually impaired. *Multimedia Tools and Applications*. 2023. URL: https://doi.org/10.1007/s11042-023-17849-7 (дата звернення: 11.06.2026).

20. Kadhim M., Oleiwi B. Blind Assistive System based on Real Time Object Recognition using Machine learning. *Engineering and Technology Journal*. 2022. Vol. 40, no. 1. P. 159–165. URL: https://doi.org/10.30684/etj.v40i1.1933 (дата звернення: 11.06.2026).

21. Rachburee N., Punlumjeak W. An assistive model of obstacle detection based on deep learning: YOLOv3 for visually impaired people. *International Journal of Electrical and Computer Engineering (IJECE)*. 2021. Vol. 11, no. 4. P. 3434. URL: https://doi.org/10.11591/ijece.v11i4.pp3434-3442 (дата звернення: 11.06.2026).

22. Schicktanz S. et al. AI-assistedethics? Considerationsof AI simulation for theethi calasses smentand desig no fassistive technologies. *FrontiersinGenetics*. 2023. Vol. 14. URL: https://doi.org/10.3389/fgene.2023.1039839 (дата звернення: 24.05.2026).

23. Recommendation on the Ethics of Artificial Intelligence. *UNESCO*. URL: https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence (дата звернення: 24.05.2026).

24. Regulation (EU) 2024/1689 laying down harmon is edruleson artificial intelligence. *EUR-Lex. Official Journal of the European Union*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (дата звернення: 24.05.2026).

25. Recommendation on the Ethics of Artificial Intelligence. *UNESCO Digital Library*. 2022. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455 (дата звернення: 24.05.2026).

26. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. EUR-Lex. *Official Journal of the European Union*. 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (дата звернення: 24.05.2026).

27. Directive (EU) 2024/2853 on liability for defective products. EUR-Lex. *Official Journal of the European Union.* 2024. URL: https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng (дата звернення: 24.05.2026).

28. The new regulation of defective products: the EU Directive. *Fieldfisher*. URL: https://www.fieldfisher.com/en/locations/espana/actualidad/the-new-regulation-of-defective-products-the-eu-di (дата звернення: 24.05.2026).

29. Buiten M. C., DeStreel A., Peitz M. The law and economics of AI liability. *Computer Law & Security Review*. 2023. Vol. 48. URL: https://doi.org/10.1016/j.clsr.2023.105794 (дата звернення: 24.05.2026).

30. Regulation (EU) 2023/988 on general product safety. EUR-Lex. *Official Journal of the European Union.* 2023. URL: https://eur-lex.europa.eu/eli/reg/2023/988/oj/eng (дата звернення: 24.05.2026).

31. Novelli C., Taddeo M., Floridi L. Account ability in artificial intelligence: what it is and how it works. *AI &Society*. 2023. URL: https://doi.org/10.1007/s00146-023-01635-y (дата звернення: 24.05.2026).

32. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free move ment of such data. EUR-Lex. *Official Journal of the European Union*. 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04/eng (дата звернення: 24.05.2026).

33. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactionson Pattern Analysis and Machine Intelligence*. 2017. Vol. 39, no. 6. P. 1137–1149. URL: https://doi.org/10.1109/TPAMI.2016.2577031 (дата звернення: 24.05.2026).

34. Ultralytics. *GitHub*. URL: https://github.com/ultralytics/ultralytics (дата звернення: 24.05.2026).

35. Howard A. et al. Searchingfor MobileNetV3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV).* 2019. P. 1314–1324. URL: https://doi.org/10.1109/ICCV.2019.00140 (дата звернення: 24.05.2026).

36. Sokolova M., Lapalme G. A system atic analysis of performance measures for classification tasks. *Information Processing & Management*. 2009. Vol. 45, no. 4. P. 427–437. URL: https://doi.org/10.1016/j.ipm.2009.03.002 (дата звернення: 24.05.2026).

37. Kumar A. et al. Navigating bey on dsight: a real-time 3D audio-enhanced object detection system for empowering visually impaired spatialaw are ness. *Signal,*

Imageand Video Processing. 2025. Vol. 19, no. 12. URL: https://doi.org/10.1007/s11760-025-04614-6 (дата звернення: 24.05.2026).

38. Hazarika A., Rahmati M. Towardsan Evolved Immersive Experience: Exploring 5G- and Beyond-Enabled Ultra-Low-Latency Communications for Augmented and Virtual Reality. *Sensors*. 2023. Vol. 23, no. 7. URL: https://doi.org/10.3390/s23073682 (дата звернення: 24.05.2026).

39. Amin M. B. et al. A visual aid system using image processing and deep learning with audio haptic feedback. *Procedia Computer Science*. 2024. Vol. 246. P. 3105–3114. URL: https://doi.org/10.1016/j.procs.2024.09.589 (дата звернення: 24.05.2026).

40. Kendall M. G. A New Measure of Rank Correlation. *Biometrika*. 1938. Vol. 30, no. 1/2. P. 81–93. URL: https://doi.org/10.1093/biomet/30.1-2.81 (дата звернення: 24.05.2026).

41. Mean opinion score (MOS) terminology: Recommendation ITU-T P.800.1. *International Telecommunication Union.* 2016. URL: https://www.itu.int/rec/T-REC-P.800.1-201607-I/en (дата звернення: 24.05.2026).

42. PyQt5. *Python Package Index*. URL: https://pypi.org/project/PyQt5/ (дата звернення: 24.05.2026).

43. OpenCV-Python Tutorials. *OpenCV documentation*. URL: https://docs.opencv.org/master/d6/d00/tutorial_py_root.html (дата звернення: 24.05.2026).

44. Ultralytics Docs. *Ultralytics*. URL: https://docs.ultralytics.com/ (дата звернення: 24.05.2026).

45. PyTorch documentation. *PyTorch*. URL: https://docs.pytorch.org/docs/main/ (дата звернення: 24.05.2026).

46. FaceRecognition documentation. *Readthe Docs*. URL: https://face-recognition.readthedocs.io/en/latest/readme.html (дата звернення: 24.05.2026).

47. Keras: Deep Learning for humans. *Keras*. URL: https://keras.io/ (дата звернення: 24.05.2026).

48. pyttsx3. *Python Package Index*. URL: https://pypi.org/project/pyttsx3/ (дата звернення: 24.05.2026).

49. QtMultimedia. *Qt Documentation*. URL: https://doc.qt.io/qt-6/qtmultimedia-index.html (дата звернення: 24.05.2026).

50. COCO128. *Ultralytics Docs*. URL: https://docs.ultralytics.com/ru/datasets/detect/coco128/ (дата звернення: 24.05.2026).

51. Eckhoff D., Schnupp J., Cassinelli A. Temporal precision and accuracy of audio-visual stimuliin mixed reality systems. *PLOS ONE*. 2024. Vol. 19, no. 1. P. e0295817. URL: https://doi.org/10.1371/journal.pone.0295817 (дата звернення: 09.06.2026).

52. Theodorou P., Tsiligkos K., Meliones A. Multi-Sensor Data Fusion Solutions for Blind and Visually Impaired: Research and Commercial Navigation Applications for Indoor and Outdoor Spaces. *Sensors*. 2023. Vol. 23, no. 12. URL: https://doi.org/10.3390/s23125411 (дата звернення: 24.05.2026).

53. Zhang Y. et al. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. ComputerVision – ECCV 2022. *Lecture Notes in Computer Science*. 2022. Vol. 13682. P. 1–21. URL: https://doi.org/10.1007/978-3-031-20047-2_1 (дата звернення: 24.05.2026).

54. Shorten C., Khoshgoftaar T. M., Furht B. A Comprehensive Survey of Image Augmentation Techniques for Deep Learning. *Pattern Recognition*. 2023. Vol. 137. URL: https://doi.org/10.1016/j.patcog.2023.109347 (дата звернення: 24.05.2026).

55. Harshitha R., Lakshmipriya B., Krishnamurthy V. TransEffiVisNet – an image captioning architecture for auditory assistance for the visually impaired. Multimedia Toolsand Applications. 2024. URL: https://doi.org/10.1007/s11042-024-20036-x (дата звернення: 24.05.2026).

56. XTTS-v2. HuggingFace. URL: https://huggingface.co/coqui/XTTS-v2 (дата звернення: 24.05.2026).

57. Мазурець О. В., Петровський С. С., Дидо Р. А. Нейромережева модель для ідентифікації особистості за зображенням обличчя у реальному часі. *Інформаційні технології і автоматизація : матеріали XVII міжнародної науково-практичної конференції*, 31 жовтня – 1 листопада 2024 р., Одеса, ОНТУ. Одеса, 2024. С. 655–658.

58. Дидо Р. А., Мазурець О. В., Кліменко В. І. Інформаційна система для нейромережевої інтерактивної ідентифікації особистості за зображенням обличчя. *Актуальні проблеми комп'ютерних наук АПКН-2024 : збірник наукових праць за матеріалами XVI Всеукраїнської науково-практичної конференції*, 15–16 листопада 2024 р. Хмельницький, 2024. С. 180–186. URL: https://kn.khmnu.edu.ua/wp-content/uploads/sites/18/apkn-2024-corpuspaper.pdf(дата звернення: 09.06.2026).

59. Дидо Р. А., Мазурець О. В. Метод ідентифікації особистості на основі розпізнавання обличчя в реальному часі для систем кібербезпеки. *Інформаційна, функційна і кібербезпека СКІФіК2024 : матеріали IV Всеукраїнської науково-технічної конференції*, 29–30 листопада 2024 р. Харків, 2024. С. 36–37. URL: http://dx.doi.org/10.13140/RG.2.2.14272.75521 (дата звернення: 09.06.2026).

60. Дидо Р., Собко О., Мазурець О. Метод формування аудіопотоку з моніторингу оточуючого середовища засобами машинного навчання. *Наука і техніка сьогодні.* 2025. № 1(42). URL: https://doi.org/10.52058/2786-6025-2025-1(42)-1148-1161 (дата звернення: 09.06.2026).

61. Mazurets O., Sobko O., Dydo R., Zalutska O., Molchanova M. Augmented reality audio stream creation using CNN: boosting inclusion and safety for visually impaired people. *CEUR Workshop Proceedings*. 2025. Vol. 4004. P. 347–361. URL: https://ceur-ws.org/Vol-4004/paper26.pdf (дата звернення: 09.06.2026).

62. А. с. № 133293 Україна. Комп'ютерна програма «Інтелектуальна інформаційна система для моніторингу середовища для осіб із порушеннями зору через аудіопотік доповненої реальності на базі машинного навчання» / О. В. Собко, Р. А. Дидо, О. В. Мазурець. *Свідоцтво про реєстрацію авторського права на твір від 31.03.2025*. URL: https://sis.nipo.gov.ua/uk/search/detail/1849025/(дата звернення: 09.06.2026).

ДОДАТКИ

Додаток А

Програмні коди

Вихідний код, використаний у дослідженні, доступний у репозиторії GitHub: https://github.com/Roman-Osinchuk/blind_assistant (дата звернення: 28.05.2026).

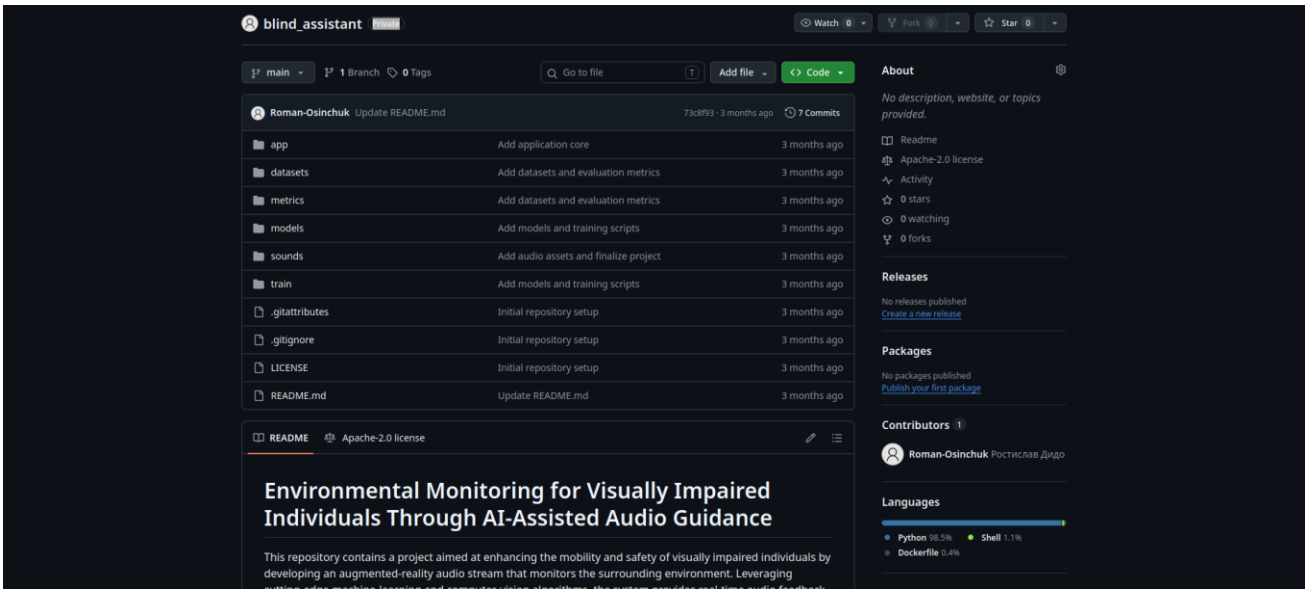


Рисунок А.1 – Головнасторінка репозиторію

Структура репозиторію наступна:

– модулі прикладного застосунку (app). Містить код користувацького застосунку, зокрема PyQt5-інтерфейс, цикл обробки відеопотоку, використання YOLO-моделей для виявлення об’єктів, розпізнавання облич та формування аудіоповідомлень для користувача;

– модулі наборів даних (datasets). Містить стиснені набори даних, які використовуються для навчання, перевірки та розширення можливостей системи, зокрема дані для розпізнавання транспортних засобів, дорожніх знаків, світлофорів, поїздів, велосипедів та окремих класів об’єктів;

– модулі оцінювання моделей (metrics). Містить інструменти для кількісної оцінки роботи моделей, зокрема конфігураційний файл набору даних, скрипт для тестування YOLO-моделей та збережені результати експериментів;

– модулі моделей (models). Містить попередньо навчені та користувацькі ваги моделей, зокрема YOLO-моделі різних версій для виявлення об’єктів, а також Keras-модель для класифікації відомих облич;

– модулі аудіосупроводу (sounds). Містить звукові ресурси, які використовуються для аудіозворотного зв’язку та створення фонових або попереджувальних звукових сигналів;

– модулі навчання (train). Містить код для навчання CNN-моделі розпізнавання облич, зокрема підготовку класів осіб, навчання моделі та збереження результатів для подальшого використання в основному застосунку.

Додаток Б

Презентаційний матеріал

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

МЕТОД ФОРМУВАННЯ АУДІОПОТОКУ ДОПОВНЕНОЇ РЕАЛЬНОСТІ ЗА ВІДЕОДАНИМИ ДЛЯ ЛЮДЕЙ ІЗ ПРОБЛЕМАМИ ЗОРУ ЗАСОБАМИ ГЛИБОКОГО НАВЧАННЯ

Виконав:
студент групи КН-22-1
Ростислав ДИДО

Керівник:
старший викладач кафедри. КН
Олена СОБКО

Актуальність

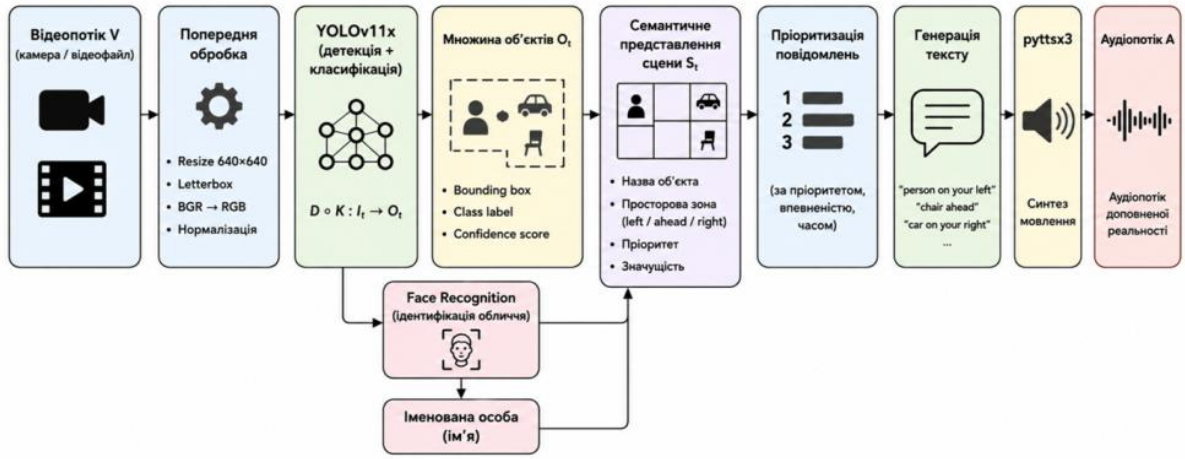
Актуальність дослідження зумовлена зростанням потреби у створенні інтелектуальних систем комп'ютерного зору та доповненої реальності, орієнтованих на підвищення рівня автономності та безпеки осіб із порушеннями зору в умовах складного міського середовища. Сучасні підходи до аналізу відеопотоку на основі глибоких згорткових нейронних мереж забезпечують високу ефективність у задачах детекції та класифікації об'єктів, однак не завжди забезпечують повноцінне семантичне представлення сцени у формі, придатній для аудіального сприйняття.

Таким чином, розроблення методу формування аудіопотоку доповненої реальності на основі глибоких згорткових нейронних мереж є актуальним завданням, що поєднує задачі комп'ютерного зору, мультимодальної обробки даних та інтерфейсів людина–комп'ютер.

Мета і задачі роботи

- **Об’єктом дослідження** є процес формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору.
- **Предметом дослідження**методи глибокого навчання для детекції та класифікації об’єктів, розпізнавання іменованих осіб та формування аудіопотоку доповненої реальності на основі відеоданих.
- **Метою роботи** є підвищення інформативності та доступності сприйняття навколишнього середовища для осіб із порушеннями зору шляхом формування аудіопотоку доповненої реальності на основі аналізу відеоданих.

Пайплайн формування аудіопотоку доповненої реальності



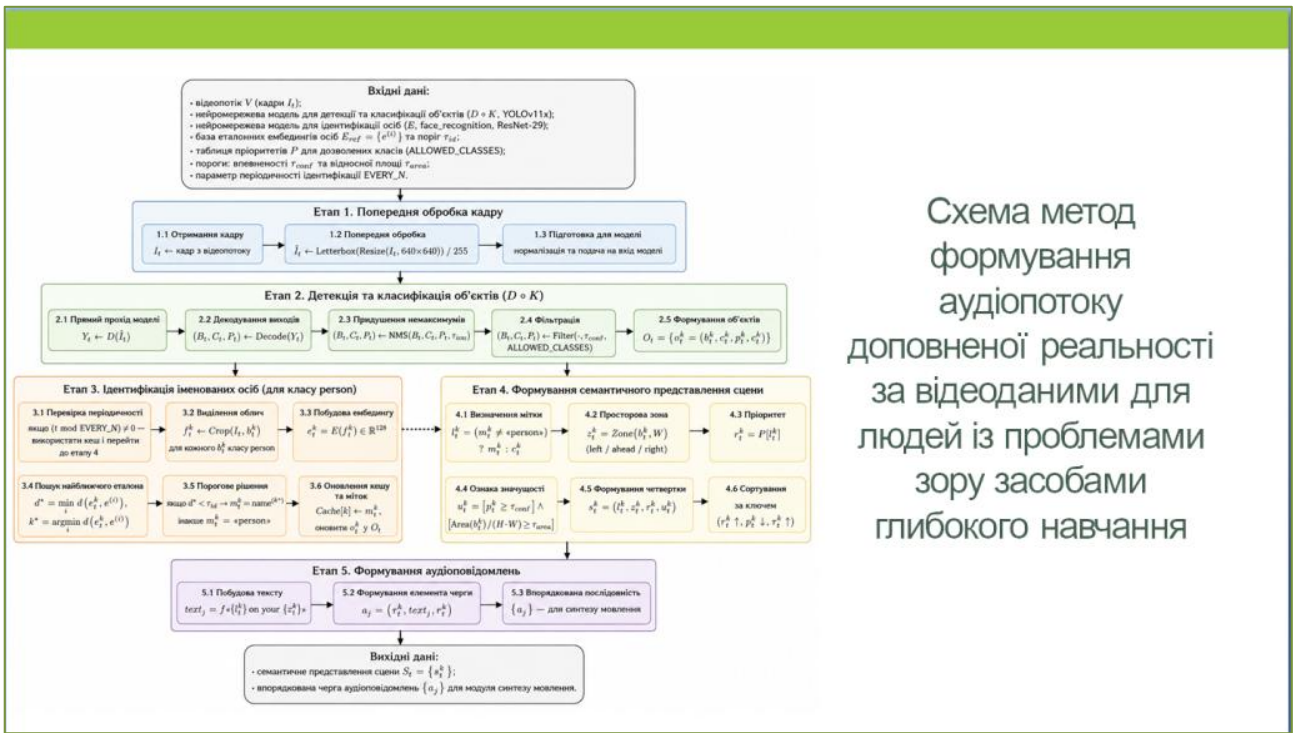
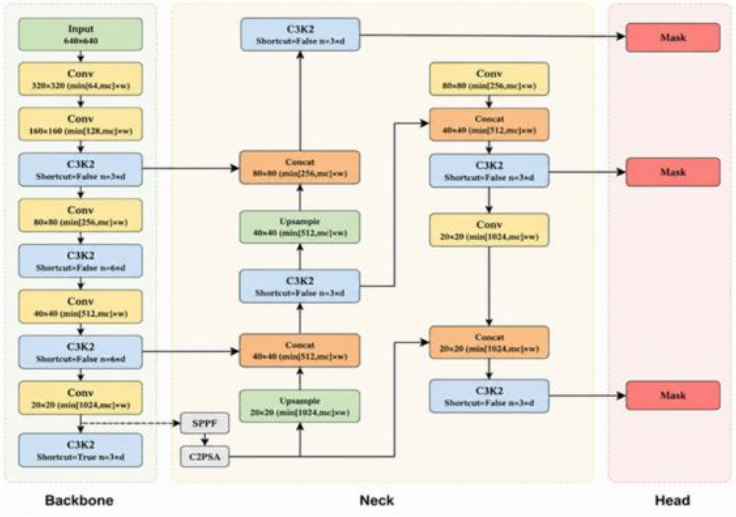


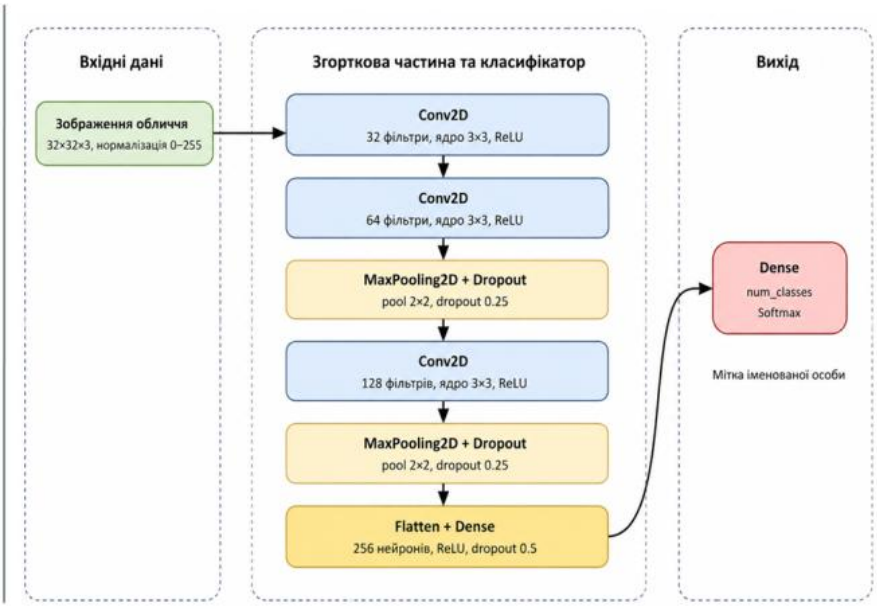
Схема метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання

Узагальнена структурна схема інтелектуальної системи

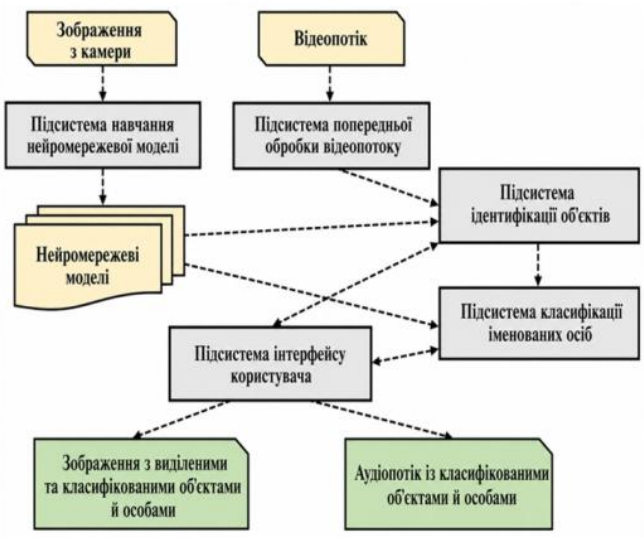
Архітектура нейромережевої моделі YOLOv11



Архітектура згорткової нейронної мережі для класифікації іменованих осіб

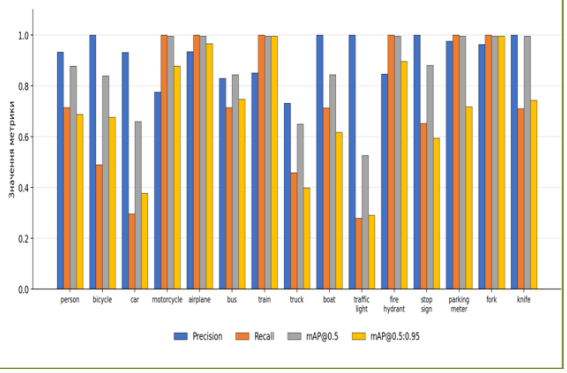


Функціональні модулі



Покласові метрики моделі YOLOv11x на тестовому наборі COCO128

Клас	Precision	Recall	mAP@0.5	mAP@0.5:0.95
person	0,933	0,715	0,877	0,688
bicycle	1,000	0,489	0,839	0,676
car	0,932	0,297	0,659	0,377
motorcycle	0,776	1,000	0,995	0,877
airplane	0,935	1,000	0,995	0,966
bus	0,829	0,714	0,843	0,747
train	0,850	1,000	0,995	0,995
truck	0,732	0,458	0,649	0,399
boat	1,000	0,713	0,843	0,617
trafficlight	1,000	0,280	0,526	0,291
firehydrant	0,846	1,000	0,995	0,896
stopsign	1,000	0,652	0,880	0,594
parkingmeter	0,975	1,000	0,995	0,718
fork	0,963	1,000	0,995	0,995
knife	1,000	0,710	0,995	0,743



Висновки

- Виконано експериментальне дослідження методу. Порівняння шести конфігурацій YOLO на тестовому наборі COCO128 показало, що найкращі усереднені метрики має модель YOLOv11x: *Accuracy* 0,741, *Precision* 0,737, *Recall* 0,659, *mAP@0.5* 0,713, *mAP@0.5:0.95* 0,549. Для *FPS* 182,66. Для CNN-класифікатора іменованих осіб установлено оптимальні гіперпараметри $batch_{size}$ = 32, *epochs* = 10, за яких отримано *Accuracy* 0,97 , *Precision* 0,97 , *Recall* 0,97 , F_1-*score* 0,95 , *verification accuracy* 0,97. На трьох тестових сценаріях якість формування аудіопотоку досягає значень: *latency* 0,42 с, *coverage* 0,93, *semantic correctness* 0,95, *priority accuracy* 0,97, *FPS* 32,8.
- Здобуті результати підтверджують, що метод забезпечує своєчасне, повне, семантично коректне та правильно впорядковане за пріоритетом представлення інформації про навколишнє середовище у формі аудіопотоку доповненої реальності.
- Практичне значення роботи полягає у можливості використання розробленого методу та інтелектуальної системи для підвищення автономності, безпеки та якості орієнтації осіб із порушеннями зору в міському середовищі та в умовах приміщення.

ДЯКУЮ ЗА УВАГУ

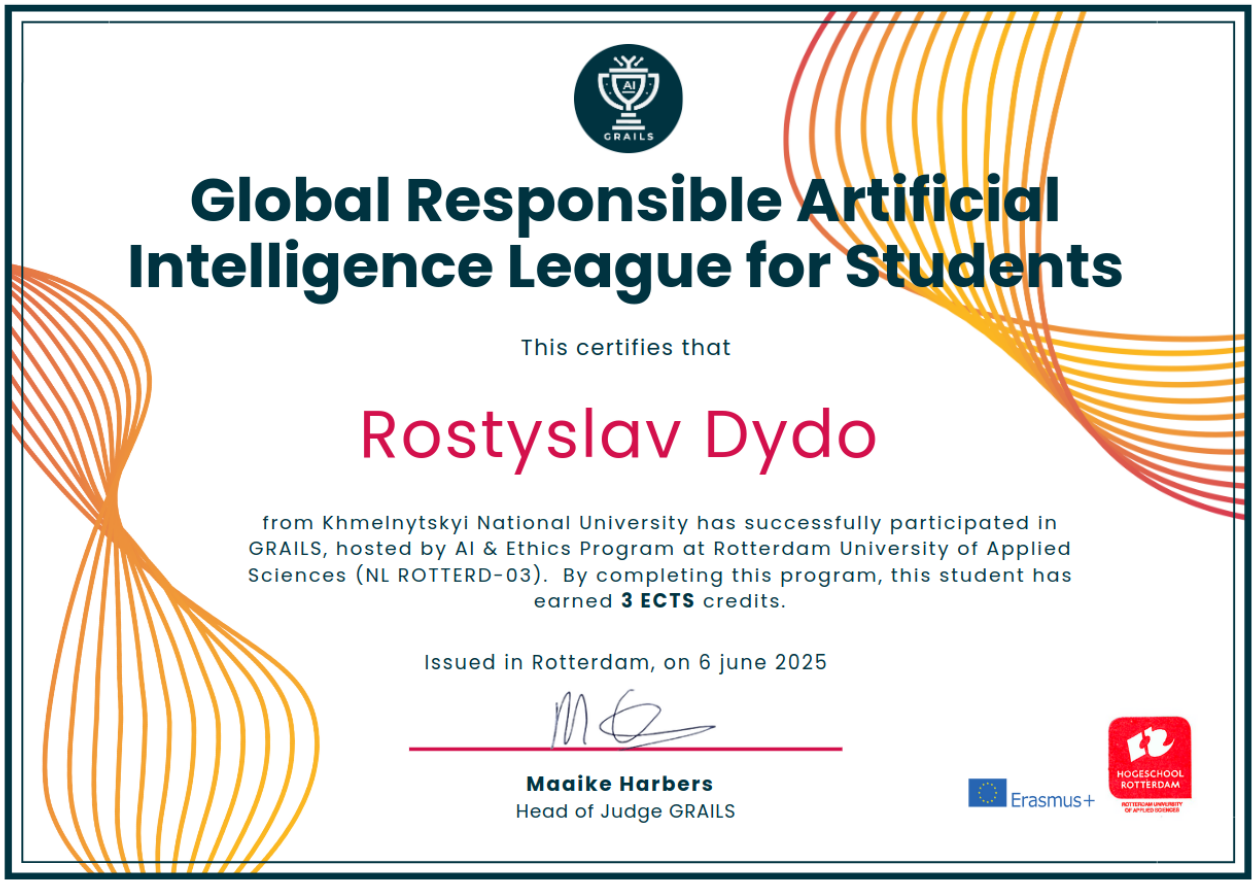
Додаток В

Сертифікат Міжнародного конкурсу студентських наукових робіт «Black Sea Science 2025»


FIELD OF «INFORMATION TECHNOLOGIES, AUTOMATION AND ROBOTICS»
IN THE INTERNATIONAL COMPETITION OF STUDENT SCIENTIFIC WORKS
«BLACK SEA SCIENCE 2025»
organized by
Odesa National University of Technology
Odesa, Ukraine
Certificate of the winner
Machine learning method for creating augmented reality audio stream to enhance the safety of people with visual impairments
authored by
Dydo Rostyslav
under the supervision of
Sobko Olena, Mazurets Oleksandr
was awarded the 1st place
Head of the Organizing Committee
Rector of Odesa National University of Technology
Larysa IVANCHENKOVA
President of Odesa National University of Technology
Bogdan IEGOROV
Vice-Rector for Scientific Work and International Relations of Odesa National University of Technology
Olga OLSHEVSKA
Head of the Jury in the field of "Information Technologies, Automation and Robotics"
Oleksandra BULGAKOVA
BSS-2025.3.45

Додаток Г

Сертифікати GRAILS 2026



Global Responsible Artificial Intelligence League for Students

This certifies that

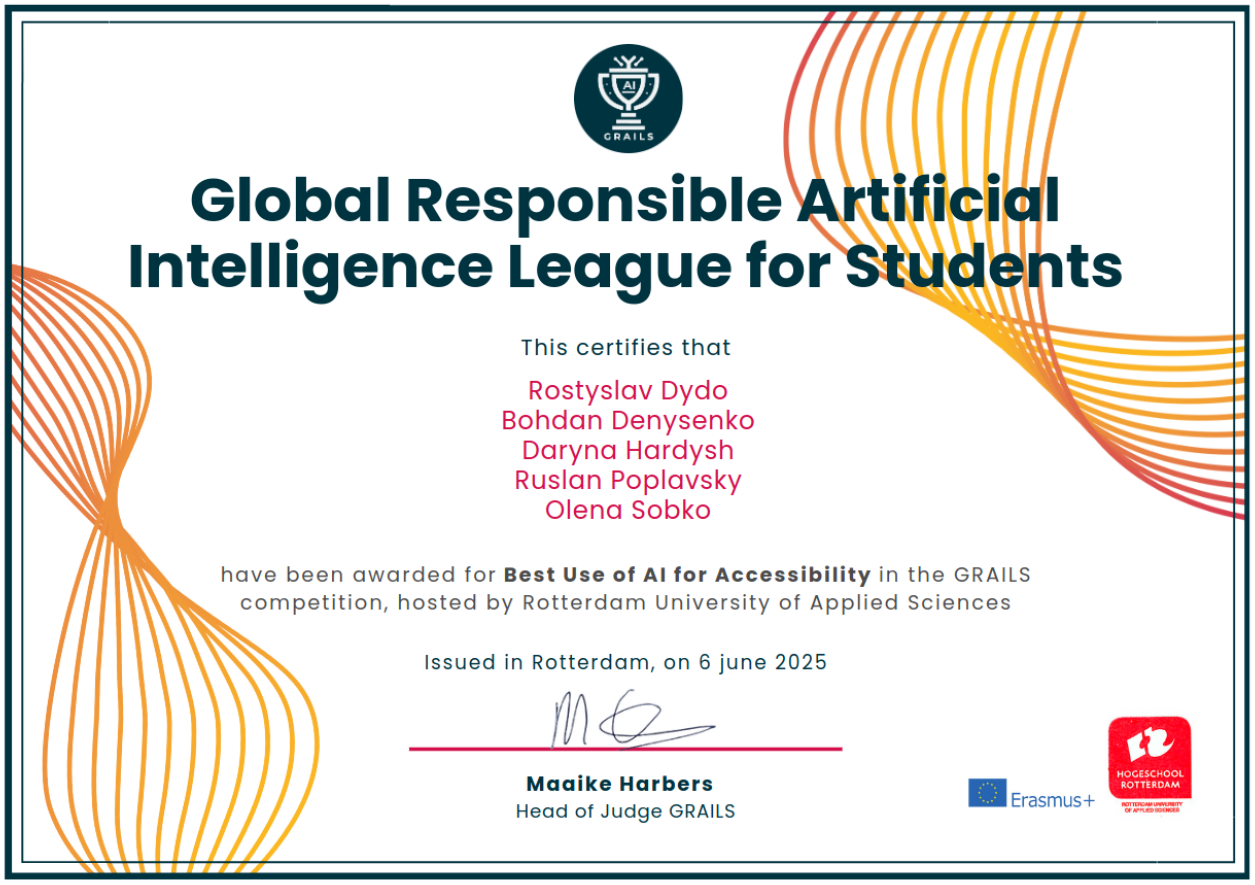
Rostyslav Dydo

from Khmelnytskyi National University has successfully participated in GRAILS, hosted by AI & Ethics Program at Rotterdam University of Applied Sciences (NL ROTTERD-03). By completing this program, this student has earned **3 ECTS** credits.

Issued in Rotterdam, on 6 june 2025

Maaike Harbers
Head of Judge GRAILS

Erasmus+



Global Responsible Artificial Intelligence League for Students

This certifies that

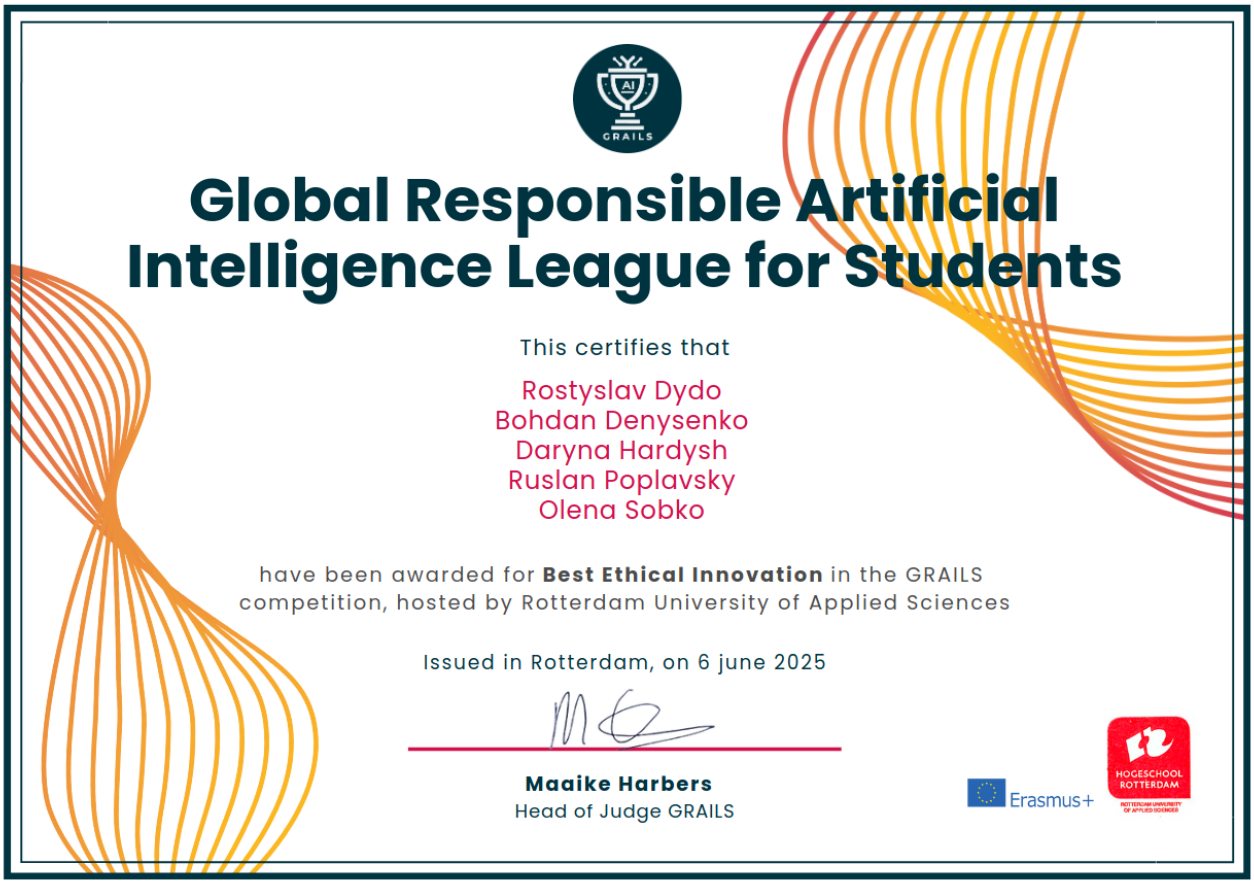
Rostyslav Dydo
Bohdan Denysenko
Daryna Hardysh
Ruslan Poplavsky
Olena Sobko

have been awarded for **Best Use of AI for Accessibility** in the GRAILS competition, hosted by Rotterdam University of Applied Sciences

Issued in Rotterdam, on 6 june 2025

Maaike Harbers
Head of Judge GRAILS

Erasmus+



Global Responsible Artificial Intelligence League for Students

This certifies that

Rostyslav Dydo
Bohdan Denysenko
Daryna Hardysh
Ruslan Poplavsky
Olena Sobko

have been awarded for **Best Ethical Innovation** in the GRAILS competition, hosted by Rotterdam University of Applied Sciences

Issued in Rotterdam, on 6 june 2025

Maaike Harbers
Head of Judge GRAILS

Erasmus+

HOGESCHOOL ROTTERDAM

Додаток Д

Свідоцтво про реєстрацію авторського права

УКРАЇНА

СВІДОЦТВО

про реєстрацію авторського права на твір

№ 133293

Комп'ютерна програма «Інтелектуальна інформаційна система для моніторингу середовища для осіб із порушеннями зору через аудіопотік доповненої реальності на базі машинного навчання»

(вид, назва твору)

Автор (співавтори) Собко Олена Віталіївна, Дидо Ростислав Андрійович, Мазурець Олександр Вікторович

(прізвище, ім'я, по батькові (за наявності), псевдонім (за наявності))

Авторські майнові права належать спільно Собко Олена Віталіївна, просп. м. Хмельницький, 29019; Дидо Ростислав Андрійович, вул. м. Стрий, Львівська обл., 82405; Мазурець Олександр Вікторович, вул. м. Хмельницький, 29027

(прізвище, ім'я, по батькові (за наявності) фізичної особи / найменування юридичної особи, адреса)

Дата реєстрації 10 лютого 2025 р.

Директор Державної організації «Український національний офіс інтелектуальної власності та інновацій» Олена ОРЛЮК

УКРНОІВІ

Wed Jun 17 08:41:54 EEST 2026, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism (http://ap.km.ua) v-16.718

Максимальне співпадіння з одним документом 2.0%

Словники перевірки: UA, US, RU. **Помилок в документах: 17%**

	Документ		Сумарний збіг по Базі Даних	
ID: 275685 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання Додано в БД: 2026-06-17 Автора: Ростислав ДИДО Керівники: Олена СОБКО Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	118650	929	4071 (3%)	55 (6%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

Протокол аналізу звіту подібності науковим керівником

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Ростислав ДИДО

Співавтор:

Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання

Науковий керівник: Олена СОБКО, к.т.н., доц. каф. КН

Підрозділ: Кафедра комп'ютерних наук

Коефіцієнт подібності 1:5.39%

Коефіцієнт подібності 2:2.41%

Мікропробіли: 0

Заміна букв: 51

Інтервали: 0

Білі знаки: 4

Дата створення звіту: 2026-06-16 19:22:14.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2026-06-17

Дата

експерт Перевезний Р.Р

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Назва кваліфікаційної роботи *Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання*
Автор *студент групи КН-22-1 Ростислав ДИДО*
Освітня програма *Комп'ютерні науки*
Рівень вищої освіти *перший (бакалаврський)*
Спеціальність *122 – Комп'ютерні науки*
Науковий керівник: *ст. викладач каф. КН, д-р філософії Олена СОБКО*

На основі аналізу кваліфікаційної роботи на дотримання вимог академічної доброчесності (у т.ч. відсутності ознак академічного плагіату) з урахуванням результатів перевірки роботи спеціалізованим програмними засобами комісія зробила такий висновок:

№	Висновок	Позначка про відповідність
1	Ознаки академічного плагіату	
1.1	Запозичення, виявлені в роботі, є законними і не є академічним плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних, якщо потрібно). Робота приймається до захисту.	***відповідає***
1.2	Виявлені запозичення не є академічним плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована.	
1.3	Виявлені запозичення не є академічним плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. Робота може бути допущена до захисту після того як буде відкоригована та доопрацьована і успішно пройде повторну перевірку на академічний плагіат.	
1.4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття текстових запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
2	Інші види порушень академічної доброчесності	***відсутні***

Підтвердження:

Запозичення, виявлені в роботі Ростислава Дидо, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти, які не мають авторства і містять поширені конструкції та загальновідомі терміни, скорочення. Рівень подібності не перевищує допустимої межі. Таким чином, робота є законною та приймається до захисту.

Обсяг запозичень, визначений системами виявлення збігів/ідентичності/схожості:
- за системою Anti-Plagiarism: 2%;
- за системою StrikePlagiarism КП1: 5.39%,.

17.06.2025

Завідувач кафедри _____ *Олександр БАРМАК*

Гарант освітньої програми _____ *Олександр МАЗУРЕЦЬ*

Керівник кваліфікаційної роботи _____ *Олена СОБКО*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *КН-22-1 Дидо Ростислава Андрійовича*

за темою *Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання*

1. Актуальність теми

Обраний напрям має не декларативну, а реальну прикладну важливість. Задача перетворення відеоданих у змістовний аудіопотік для людей із порушеннями зору потребує автоматизованих підходів, оскільки ручне опрацювання таких випадків є повільним, неоднорідним і залежним від людського чинника. Під час роботи над темою важливо було врахувати не тільки алгоритмічний аспект, а й специфіку даних, з якими працює система. Саме тому дослідження має достатню актуальність для бакалаврського рівня і відповідає сучасним тенденціям розвитку комп'ютерних наук.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

За описом предметної області спеціальності 122 Комп'ютерні науки об'єктом роботи є процес формування доступного аудіосупроводу за відеопотоком. Метою кваліфікаційної роботи бакалавра є підвищення якості формування аудіопотоку доповненої реальності для осіб із порушеннями зору, що полягає у підвищенні точності детекції та класифікації об'єктів, точності розпізнавання іменованих осіб, своєчасності формування контекстно значущих аудіоповідомлень та повноти передачі інформації про навколишнє середовище користувачу.

3. Професійні та особистісні якості бакалавра

Робота виконувалася без формального ставлення до поставлених завдань. Студент вміє працювати з фаховою інформацією і достатньо впевнено пояснює логіку власних рішень. Автор працював послідовно, без формального ставлення до окремих етапів дослідження. Позитивним є те, що здобувач не обмежився поверховим описом технологій, а намагався пояснити логіку їх застосування в межах власної задачі.

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Ступінь самостійності здобувача під час виконання кваліфікаційної роботи можна оцінити позитивно. Автор самостійно здійснив аналіз предметної області,

обґрунтував вибір методів дослідження, виконав програмну реалізацію запропонованого рішення та провів експериментальну перевірку отриманих результатів.

5. Ступінь оволодіння методами дослідження

Під час виконання роботи студент не лише застосував готові інструменти, а й пояснив логіку їх вибору. Ступінь оволодіння методами дослідження є належним: у роботі поєднано аналіз літератури, формалізацію задачі, програмну реалізацію та експериментальну перевірку.

6. Повнота та якість розкриття теми роботи

Тема розкрита достатньо повно. У роботі є аналіз предметної області, обґрунтування методу, опис програмної реалізації та оцінювання результатів. Матеріал не виходить за межі теми й водночас не зводиться до поверхового огляду. Якість розкриття теми проявляється в тому, що автор пов'язує теоретичні положення з практичною реалізацією. Робота не обмежується загальними міркуваннями про візуальних даних і методів комп'ютерного зору, а демонструє конкретний шлях розв'язання поставленої задачі.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Структура роботи зрозуміла: від постановки задачі автор переходить до методу, а потім до реалізації та результатів. Аргументація достатня для бакалаврського рівня. Робота написана зрозумілою академічною мовою, без надмірного ускладнення там, де це не потрібно.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Результати можуть бути корисними не лише як навчальний проєкт, а й як база для подальшого вдосконалення методу, розширення набору даних і перевірки в реальніших умовах. Результати роботи мають прикладний потенціал, оскільки можуть бути адаптовані до реальних сценаріїв використання. Водночас для промислового застосування потрібне подальше розширення даних, тестування та уточнення параметрів системи.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

З огляду на зміст, самостійність виконання та практичну спрямованість результатів роботу можна рекомендувати до захисту. Робота заслуговує на оцінку « відмінно *».*

Керівник _____ *ст. викладач каф. КН, д-р філософії Олена СОБКО*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-22-1 Дидо Ростислава Андрійовича*
за темою: *Метод формування аудіопотоку доповненої реальності за відеоданими для людей із проблемами зору засобами глибокого навчання*

1. Актуальність обраної теми

Актуальність роботи визначається прикладною значущістю задачі перетворення відеоданих у змістовний аудіопотік для людей із порушеннями зору. У сучасних умовах такі задачі дедалі частіше потребують не ручного аналізу, а відтворюваних інтелектуальних методів. Окремо слід підкреслити, що тема має міждисциплінарний характер: у ній поєднуються методи комп'ютерних наук, аналіз даних і практична потреба конкретної сфери застосування. Це робить дослідження не формальним, а таким, що має потенціал подальшого розвитку.

2. Повнота розкриття мети та завдань роботи

Завдання роботи в цілому виконані. Повнота розкриття проявляється у переході від аналізу джерел до формалізації методу та оцінювання результатів. Така логіка дозволяє побачити, що виконання роботи не обмежувалося компіляцією відомих положень, а передбачало власне опрацювання матеріалу. Важливо, що виконання завдань має послідовний характер: кожний наступний етап спирається на попередній. Така побудова свідчить про розуміння логіки дослідження, а не лише про формальне виконання пунктів завдань роботи.

3. Зміст кожного розділу роботи

У роботі послідовно розглянуто предметну область, модельну частину і програмну реалізацію. Така побудова є зручною для оцінювання, оскільки видно, як вихідна проблема переходить у конкретне технічне рішення. Зміст розділів подано достатньо розгорнуто для бакалаврського рівня. Позитивним є те, що робота містить не тільки загальний огляд предметної області, а й пояснення особливостей даних, етапів обробки та критеріїв оцінювання. У роботі враховано, що для візуальних даних недостатньо просто застосувати готовий алгоритм: потрібно пояснити, які обмеження має обраний підхід і як вони можуть вплинути на результати.

4. Оцінка розробленої інформаційної системи, її практична цінність

Практична цінність полягає в можливості використання результатів для підвищення автономності користувачів і доступності цифрових AR-сервісів. Розробка не виглядає відірваною від реального застосування: вона орієнтована на обробку конкретних даних і отримання інтерпретованого результату. Розроблена система має прикладний характер, оскільки орієнтована на обробку конкретних типів даних і отримання результату, придатного для інтерпретації.

5. Якість оформлення кваліфікаційної роботи бакалавра

Матеріал подано достатньо охайно й послідовно, без різких переходів між теоретичними та практичними фрагментами. Оформлення підтримує загальну логіку

роботи: заголовки, пояснення, результати та висновки розміщено у зрозумілій послідовності. Такий рівень подання полегшує оцінювання і створює позитивне враження від роботи.

6. Недоліки кваліфікаційної роботи бакалавра

До зауважень можна віднести певну стисліcть у поясненні параметрів експериментальної перевірки. Варто врахувати, що якість вхідних зображень, варіативність сцен і коректність розмітки суттєво впливають на результат, тому ширше тестування могло б зробити висновки ще переконливішими. Проте для бакалаврського рівня представлений обсяг дослідження є достатнім.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота

Кваліфікаційна робота відповідає вимогам до бакалаврських робіт за спеціальністю 122 Комп'ютерні науки, може бути рекомендована до захисту та заслуговує на оцінку « відмінно *».*

Рецензент к.т.н, доц. Калустян М.В. _____