https://www.sworldjournal.com/index.php/swj/article/view/swj34-01-113
DOI: 10.30888/2663-5712.2025-34-01-113

УДК 004.89:004.912

DEVELOPMENT OF A MULTIMODAL RECOGNITION SYSTEM AND TARGET CLASSIFICATION BASED ON ARTIFICIAL INTELLIGENCE

РОЗРОБЛЕННЯ СИСТЕМИ МУЛЬТИМОДАЛЬНОГО РОЗПІЗНАВАННЯ ТА КЛАСИФІКАЦІЇ ЦІЛЕЙ НА ОСНОВІ ШТУЧНОГО ІНТЕЛЕКТУ

Nikolyuk P. K./Ніколюк П.К.
d. phys.-mat. s., prof. / д. фіз.-мат. н., проф.
ORCID: 0000-0002-0286-297X
Sapozhnikova V. Y./ Сапожнікова В. Є.
student/студент
Chemes V. S./Чемес В. С.
student/студент
Vasil Stus' Donetsk National University, Vinnytsia, 600-richchia, 21, 21021
Донецький національний університет імені Василя Стуса, Вінниця, 600-річчя, 21, 21021

Анотація. *У роботі розглянуто проблему розроблення системи мультимодального розпізнавання та класифікації цілей на основі штучного інтелекту. Актуальність теми зумовлена необхідністю підвищення надійності автоматичного виявлення об'єктів в умовах низької видимості, зміни освітлення, завад та варіативності атмосферних умов. Для цього застосовуються мультимодальні підходи, які поєднують інформацію з видимого (RGB) та теплового (IR/TIR) діапазонів, що дає змогу суттєво підвищити стійкість та точність моделей. Проаналізовано три репрезентативні датасети: MFNet/ir seg, PST900 RGB-T та SemanticRT, які містять узгоджені RGB та теплові зображення для задач сегментації, детекції та класифікації об'єктів. Проведено огляд сучасних підходів до мультимодального глибинного навчання на основі штучного інтелекту, включно з методами раннього, середнього та пізнього злиття (Early, Mid, Late Fusion), а також моделей із контурно-орієнтованим наглядом (edge supervision). На основі аналізу визначено, які архітектури є найбільш ефективними для класифікації цілей у різних умовах – від денних і нічних періодів спостереження до ситуацій зі слабким освітленням та підвищеною роллю теплових даних. Результати роботи дозволяють сформувати системне уявлення про можливості штучного інтелекту в мультимодальному розпізнаванні, визначити оптимальні архітектури та моделі для продовження експериментальних досліджень.*

Ключові слова: *мультимодальне розпізнавання, класифікація цілей, штучний інтелект, RGB-T дані, глибинне навчання, контурно-орієнтоване навчання, сегментація.*

Вступ.

Сучасні системи автоматичного розпізнавання об'єктів все частіше застосовуються у складних умовах, таких як зміна освітлення, низька видимість, наявність тіней або освітлених місць, а також при наявності часткових перекрить об'єктів. Традиційні методи обробки зображень на основі однієї модальності, зокрема лише видимого спектра (RGB), часто не забезпечують достатньої надійності та точності в таких умовах. Відтак, використання мультимодальних даних, таких як комбінація RGB та теплових (IR) зображень, дозволяє суттєво

покращити ефективність розпізнавання об'єктів і класифікації цілей, забезпечуючи більш стійку роботу системи у різноманітних умовах [1, 2].

Застосування штучного інтелекту та глибинного навчання у цій сфері відкриває нові можливості для розробки автоматизованих систем спостереження, безпілотних платформ та інтелектуальних систем безпеки. Наприклад, дослідження Ha et al. (2017) представило MFNet, який поєднує RGB та теплову модальності для сегментації та класифікації об'єктів у міських умовах [3]. Аналогічно, PST900 – це RGB-T датасет, призначений для автоматичного розпізнавання людей у складних умовах підземних шахт, де освітлення обмежене, присутні тіні та високий контраст. Використання цього датасету дозволяє тестувати моделі мультимодального розпізнавання в умовах, близьких до реальних промислових сценаріїв [4]. У свою чергу SemanticRT забезпечує великий обсяг даних для тестування сучасних моделей мультимодальної сегментації та класифікації [5]. Результати цих робіт показують, що інтеграція кількох модальностей значно підвищує точність та стабільність системи, особливо у складних і нічних умовах.

Зважаючи на актуальність проблеми, наукова спільнота звертає увагу на розробку ефективних підходів до мультимодального злиття даних (Early, Mid та Late Fusion), а також на моделі з контурно-орієнтованим навчанням (edge supervision), що дозволяє покращити локалізацію об'єктів на зображенні. Останні публікації свідчать, що використання таких методів у поєднанні з сучасними мережевими архітектурами такими як RTFNet та ECM, дає значне підвищення точності розпізнавання [3, 4, 5].

Таким чином, проблема створення надійної та ефективної системи мультимодального розпізнавання та класифікації цілей має не тільки наукове, а й практичне значення. Вона є критично важливою для задач охорони, безпеки, автономного керування безпілотними літальними апаратами та моніторингу міських середовищ, де традиційні методи не справляються зі складними умовами освітлення і погодними факторами.

Метою роботи є розроблення системи мультимодального розпізнавання та

класифікації цілей на основі штучного інтелекту, яка здатна ефективно поєднувати інформацію з RGB та теплових зображень для підвищення точності класифікації об’єктів у різних умовах сцени.

Для досягнення цієї мети в роботі передбачено комплексне дослідження існуючих датасетів, таких як MFNet / ir_seg, PST900 та SemanticRT, з метою оцінки їх придатності для мультимодальних задач сегментації та класифікації об’єктів. Крім того, планується аналіз і порівняння сучасних архітектур моделей з різними підходами до злиття модальностей – Early, Mid та Late Fusion, а також розробка методів з використанням контурно-орієнтованого навчання, що дозволяють покращити локалізацію об’єктів на зображеннях. На основі отриманих результатів передбачається визначення оптимальної архітектури, яка забезпечить максимальну точність сегментації та класифікації цілей у складних умовах сцени, включно з низькою освітленістю та високим контрастом між об’єктами та фоном.

Огляд літератури.

Мультимодальне розпізнавання об’єктів на основі RGB та теплових зображень є одним із пріоритетних напрямів сучасного штучного інтелекту. Застосування лише однієї модальності часто не забезпечує достатньої надійності в умовах низької освітленості, високого контрасту чи часткових перекрить об’єктів [1, 2]. Використання мультимодальних даних дозволяє підвищити точність сегментації та класифікації, що має важливе практичне значення для безпілотних систем, промислового контролю та інтелектуальних систем спостереження.

Датасет ir_seg [3] є однією з перших структурованих вибірок RGB–T зображень, які застосовуються для навчання й оцінювання моделей сегментації у мультимодальних умовах. На основі цього датасету Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku та T. Harada (2017) розробили модель MFNet – а першу широковідому архітектуру для RGB–T семантичної сегментації, що використовує раннє злиття (Early Fusion) [3]. У ній теплові й видимі канали об’єднуються на початкових етапах обробки, що дало змогу експериментально

продемонструвати потенціал мультимодальності у складних умовах освітлення. Перевагою такого підходу є здатність мережі навчатися спільним ознакам різних модальностей, що підвищує точність класифікації об'єктів в умовах слабкого освітлення. Водночас раннє злиття обмежує гнучкість моделі при роботі із сильно різнорідними даними, що призвело до появи Mid Fusion та attention-based архітектур.

Датасет PST900 орієнтований на автоматичне розпізнавання людей у складних промислових умовах підземних шахт. Liu et al. (2016) показали, що інтеграція теплових даних значно покращує детекцію у темряві та складних сценах, де RGB-зображення самі по собі є недостатньо інформативними [4]. Це підкреслює практичну цінність мультимодальних систем для реальних умов експлуатації.

SemanticRT [5] надає великий набір семантично анотованих RGB-T зображень. Моделі, протестовані на ньому, включають ECM та GRA-NCD, які застосовують контурно-орієнтоване навчання (edge supervision) для покращення локалізації об'єктів [5]. Це дозволяє точніше визначати межі об'єктів навіть у випадках часткового перекриття або низького контрасту, що особливо важливо для промислових і автономних систем.

Підхід Mid Fusion / Attention Fusion, представлений у RTFNet [6] та MMTM [7], інтегрує RGB та теплові канали на проміжних рівнях нейронної мережі із застосуванням механізмів уваги. Це дає змогу виділяти найбільш релевантні ознаки кожної модальності та підвищувати точність сегментації в умовах шуму й складних сцен. Подібні концепції реалізовані в RSFNet [8], де впроваджено залишкове просторове злиття модальностей, а також у CAFNet [9], який виконує адаптивне cross-modal злиття через канал-просторову увагу. Інший підхід – CSRPNet [10] – використовує поширення просторових та канальних зв'язків між RGB і тепловими ознаками, зберігаючи корисну інформацію кожної модальності до інтеграції на рівні декодера.

Ранні, проміжні та пізні стратегії злиття модальностей забезпечують різні способи інтеграції RGB і теплових даних. Раннє злиття використовує спільний

вхід і має найпростішу архітектуру, проте гірше пристосовується до значних відмінностей між модальностями. Проміжні підходи об'єднують ознаки на глибинних рівнях і дають змогу моделі вибірково підсилювати корисні компоненти сигналу, що підвищує стійкість у складних сценах. Пізнє злиття передбачає незалежну обробку кожної модальності з подальшим комбінуванням результатів, підвищуючи стійкість до шуму та деградації окремих каналів [11, 12].

Крім того, низка сучасних робіт [13–14] демонструє ефективність контурно-орієнтованих методів у мультимодальних сегментаційних моделях. Наприклад, CMX [13] впроваджує cross-modal fusion із трансформерною увагою, що покращує відтворення контурів і локальних деталей об'єктів. У Edge-Supervised Attention-Aware Fusion Network [14] застосовується edge-aware модуль, який покращує точність меж шляхом спрямованого контролю за контурною інформацією. Ці роботи підтверджують, що інтеграція контурів є важливою складовою високоточної RGB-T сегментації.

Отже, мультимодальні RGB-T методи суттєво перевершують одно-модальні рішення. Особливу увагу слід приділяти механізмам адаптивного злиття модальностей і використанню контурної інформації, які істотно впливають на точність сегментації та класифікації.

Опис датасетів.

IR Seg Dataset [3] – це спеціалізований датасет, створений для задач сегментації об'єктів у зображеннях, зокрема для сценаріїв денного та нічного часу доби. Основною метою його розробки є підтримка досліджень у галузі комп'ютерного зору, зокрема для навчання моделей, здатних точно ідентифікувати та сегментувати різні типи об'єктів на дорожніх сценах, включаючи транспортні засоби, пішоходів та елементи дорожньої інфраструктури. Датасет орієнтований на підвищення точності алгоритмів семантичної сегментації у різних умовах освітлення, що є критично важливим для застосувань у системах автономного водіння та моніторингу дорожнього руху.

У датасеті містяться оригінальні файли анотацій у форматі JSON, де кожен об'єкт описується полігоном, тобто послідовністю точок з координатами x та y. Зображення загалом налічують 1569 одиниць, з яких 820 віднесено до денного часу, а 749 – до нічного. Анотації також перетворено на зображення істини (ground truth), де пікселі відповідають певним класам: від 0 для невизначених областей до 8 для транспортних засобів, пішоходів, велосипедів, доріг, стоячих автомобілів, відбійників, конусів і лежачих поліцейських.

Датасет включає перелік зображень без анотацій, що дозволяє уникнути їх використання під час навчання. Для збільшення навчальної вибірки перед тренуванням застосовується скрипт для генерації дзеркально відображених версій зображень. Організація даних для навчання та тестування реалізована через текстові списки, що розділяють набори для тренування, валідації, тестування, а також окремо для денних і нічних умов, що забезпечує контроль над експериментальними сценаріями.

PST900 [4] – це науково орієнтований датасет, розроблений для задач семантичної сегментації в мультисенсорному середовищі, що поєднує дані видимого спектру (RGB) та теплові (Thermal) зображення. Основною метою створення датасету є забезпечення високоякісного ресурсу для досліджень у сфері комп'ютерного зору та автономних систем, орієнтованих на складні підземні або слабо освітлені середовища. Датасет містить матеріали, зняті у шахтах та інших підземних локаціях, що дозволяє моделювати реальні сценарії роботи в обмежених умовах видимості та підвищує актуальність досліджень для застосувань у промислових, рятувальних і робототехнічних контекстах.

Датасет містить 894 строго синхронізовані пари RGB- і теплових зображень, які пройшли калібрування для забезпечення точного вирівнювання пікселів між обома модальностями. Кожне зображення супроводжується анотаціями на рівні пікселів, що охоплюють чотири класи об'єктів, окрім фону: вогнегасник, рюкзак, дриль і людина у тепловому спектрі. Таке детальне розмічення дозволяє дослідникам тренувати моделі, здатні ефективно інтегрувати інформацію з різних сенсорів для підвищення точності сегментації.

Особливістю PST900 є наявність як оброблених теплових зображень у 8-бітному форматі, так і необроблених даних у 16-бітному форматі, що дає змогу аналізувати вплив якості сенсорних даних на результати моделей. Датасет пройшов ретельне калібрування, включно з використанням портативної пасивної цілі, що забезпечує точне вирівнювання RGB- та теплових кадрів. Крім того, для роботи з тепловими зображеннями передбачені методи заповнення «прогалин», які виникають через паралакс. Це дозволяє отримувати цілісні теплові карти для навчання нейронних мереж.

SemanticRT [5] – це науково орієнтований датасет, розроблений для задач мультисенсорної семантичної сегментації, який містить синхронізовані RGB-зображення та теплові кадри. Основною метою створення датасету є забезпечення дослідників великим і якісним ресурсом для тренування моделей, здатних ефективно працювати в умовах слабкого або мінливого освітлення, де використання лише RGB-зображень є недостатнім.

Датасет включає 11 371 синхронізовану пару RGB- і теплових зображень, кожне з яких має анотації на рівні пікселів. Розмітка охоплює різні класи об’єктів, включно з фоном, людьми, автомобілями, велосипедами, мотоциклами, велосипедистами, мотоциклістами, світлофорами, коробками, стовпами та елементами дороги, такими як повороти. Завдяки такому різноманіттю класів датасет дозволяє моделювати як статичні, так і рухомі об’єкти в урбаністичних та промислових сценах, забезпечуючи широкий спектр сценаріїв для навчання моделей.

SemanticRT акцентує увагу на складних умовах освітлення, включно з нічними або темними середовищами, де теплові зображення доповнюють RGB, забезпечуючи надійне виявлення об’єктів. Дані організовані у тренувальний, валідаційний та тестовий набори. При цьому тестовий набір додатково поділено за умовами освітлення для оцінки продуктивності моделей у денних і нічних сценаріях.

Для полегшення роботи з анотаціями використовується колірна карта, де кожен клас має свій RGB-колір, що дозволяє легко візуалізувати семантичні

мітки. Таким чином, SemanticRT є масштабним і високоякісним ресурсом для досліджень у сфері мультисенсорної семантичної сегментації, особливо у складних умовах освітлення та урбаністичних середовищах.

Опис моделей мультимодального розпізнавання.

Моделі мультимодальної сегментації, що інтегрують RGB та теплову інформацію, реалізують різні підходи до комбінування ознак, і саме спосіб злиття визначає їхню ефективність у складних умовах освітлення. Однією з базових архітектур є MFNet, що була запропонована На et al. [3] як компактна модель раннього злиття для задач реального часу. Її структура включає два незалежні енкодери MiniNet для RGB та теплового каналів, після чого відбувається пряме конкатенування проміжних ознак. Такий підхід мінімізує обчислювальні витрати, зберігаючи достатню гнучкість для роботи в умовах низької освітленості, проте залишається менш чутливим до шумів і різнорідності модальностей, що обмежує якість високорівневого семантичного узагальнення. Попри це, MFNet залишається ключовою відправною точкою для багатьох робіт з раннього злиття і застосовується як базова модель для досліджень на IR Seg Dataset.

Моделі, розроблені для датасету PST900, демонструють більш складну структуру і тяжіють до комбінованого (переважно пізнього) злиття. Базова архітектура PST900 використовує двопотоковий підхід, у якому RGB-стрім на основі ResNet-18 функціонує як окремий сегментаційний модуль з U-подібним декодером, що формує початкові карти ймовірностей, тоді як теплові дані та попередні RGB-прогнози подаються до другого потоку на основі ERFNet. Така організація дозволяє реалізувати поетапне уточнення сегментації: RGB відповідає за деталізацію структури сцени, тоді як тепловий канал усуває неоднозначності та підсилює контури об’єктів у темних або зашумлених регіонах. Подібний принцип пізнього злиття також розвивається у роботах Park et al. [11] та Valada et al. [12], де багатогілкові архітектури підвищують стійкість до міжмодальних відмінностей.

На відміну від попередніх моделей, архітектура SemanticRT із вбудованим

Edge-Constrained Method тяжіє до проміжного злиття з явною контурною регуляризацією. Її двопотокові енкодери на основі модифікованих ResNet-блоків отримують ознаки з обох модальностей окремо, після чого здійснюється поетапне багаторівневе злиття за допомогою гібридних блоків, що інтегрують як низькорівневі, так і високорівневі ознаки. Ключовою інновацією моделі є використання явного контурного сигналу, який організовує процес злиття та підвищує точність локалізації об'єктів. Підхід ґрунтується на ідеях boundary-aware сегментації, характерних для GRA-NCD [5] та edge-aware RGB-T моделей, таких як метод Chen et al. [14]. Використання глобальних та семантичних модулів (GIM та SIM) додатково розширює контекст, що є критично важливим у темних сценаріях із великою кількістю дрібних або частково закритих об'єктів.

Виклад основного матеріалу дослідження.

Порівняння трьох мультимодальних моделей семантичної сегментації – MFNet, PST900 та SemanticRT – вимагало ретельної підготовки даних і стандартизації умов тестування, що дозволило забезпечити коректність подальшого аналізу. Кожна з вибраних моделей працює з двома різними модальностями: видимим RGB-спектром та тепловим інфрачервоним каналом, які у поєднанні формують більш повне представлення сцени. Оскільки обидва типи зображень відрізняються не лише за діапазоном значень, а й за геометричними характеристиками, процес підготовки був критично важливим етапом.

На початковій стадії RGB- і thermal-зображення проходили нормалізацію, яка приводила до стабільного діапазону значень та дозволяла уникнути різких перепадів у яскравості або контрастності між кадрами різних датасетів. Після цього виконувалося їх просторове вирівнювання. Оскільки камери RGB і теплового спектра мають різні кути огляду й оптичні спотворення, то навіть незначні розбіжності в геометрії могли призвести до неправильного зіставлення відповідних пікселів. Тому застосовувалося корекція перспективи та узгодження роздільності, що забезпечувало відповідність структури сцени в обох каналах. Для датасету IRSeg додатково були згенеровані оновлені теплові мапи, необхідні

для коректного подання thermal-каналу та забезпечення сумісності вхідних даних між усіма трьома моделями.

Створення цих мап дозволило вирівняти якість та структуру теплової інформації з іншими використовуваними датасетами й гарантувало коректність подальшого об’єднання модальностей під час інференсу. Усі три моделі тестувалися виключно у режимі прямого передбачення, без повторного навчання чи зміни вагових коефіцієнтів. Такий підхід дозволив об’єктивно оцінити їхню реальну продуктивність у відкритих реалізаціях. Архітектурно MFNet застосовує двогілкове кодування RGB і теплових ознак, а потім виконує їх раннє злиття. Це дає змогу об’єднати інформацію з двох спектрів на ранньому етапі і забезпечити модель багатшим набором ознак для подальшої сегментації. PST900 має більш глибоку структуру та реалізує каскадне об’єднання ознак у проміжних шарах, що дозволяє їй краще уловлювати складні закономірності сцени. SemanticRT розроблена як легковагова модель із вбудованим механізмом уточнення контурів, який поліпшує структурну цілісність сегментованих областей навіть у випадках з низькою якістю сигналу.

Візуальні результати для трьох моделей були представлені в уніфікованій структурі, що включала чотири взаємопов’язані зображення: оригінальний RGB-кадр, тепловий канал, карту семантичної сегментації та зображення з накладеними контурами об’єктів. Такий формат дозволив всебічно оцінити поведінку моделей – від загальної коректності класифікації до точності відтворення меж і збереження дрібних структур. Одним із важливих аспектів був аналіз геометричної відповідності між сегментованими об’єктами та їх реальними контурами, оскільки побудова рамки напряму показувала якість роботи моделі на рівні піксельної локалізації.

Візуальні результати роботи моделей наведені на рисунках 1–3.

Порівняння моделей виконується за основною метрикою якості семантичної сегментації mIoU, оскільки саме вона дає змогу найбільш об'єктивно оцінити точність розмежування об’єктів на багатоспектральних зображеннях. Додатково враховується найкращий досягнутий рівень точності (SOTA mIoU) на кожному

датасеті, що дозволяє визначити актуальний стан розвитку методів і кількісно оцінити відставання порівнюваних моделей від сучасних передових підходів. Крім того, для аналізу враховується специфіка кожного датасету, така як кількість класів, тип середовища, наявність складних умов спостереження (темрява, задимлення, перешкоди), різниця температурних характеристик у сцені. Ці фактори прямо впливають на досяжну точність і стабільність алгоритмів. Результати порівняння представлені у таблиці 1.

Рис. 1. Результати MFNet у форматі 2×2

Рис. 2. Результати PST900 у форматі 2×2

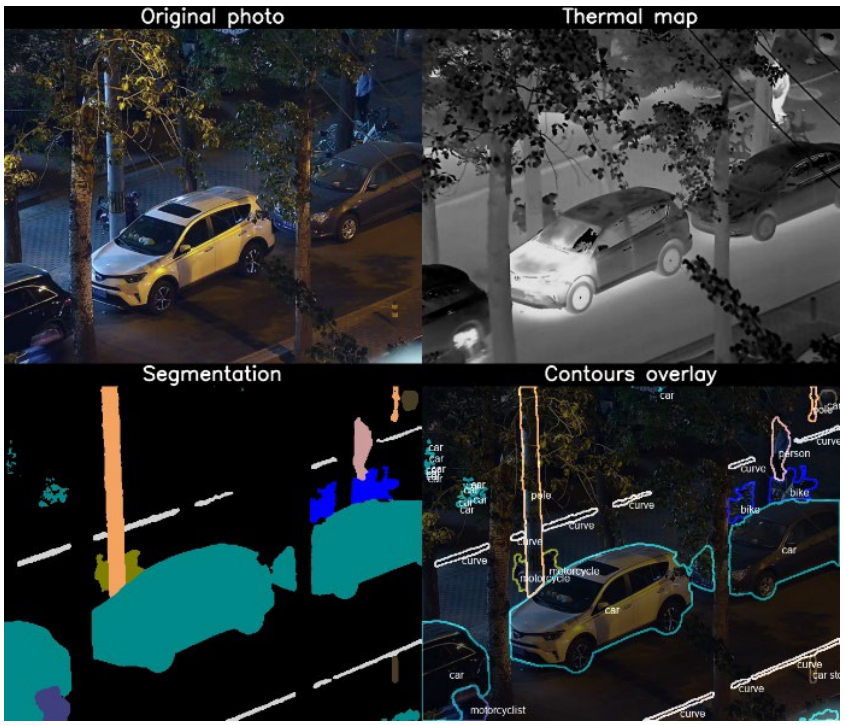

Рис. 3. Результати SemanticRT у форматі 2×2

Таблиця 1 - Порівняння продуктивності моделей на датасетах IR Seg Dataset, PST900 та SemanticRT

Модель	Датасет	mIoU(%)	SOTA mIoU	SOTA Модель
MFNet	IR Seg Dataset	53,3	57.9	SGFN (MiT-B4)
PST900	PST900	77.5	85.6	SHIFNet
ECM	SemanticRT	79.26	84.47	ECM

На основі отриманих результатів можна зробити кілька висновків. По-перше, моделі для PST900 показують найвищий базовий рівень mIoU серед розглянутих датасетів – 77.5%. Поточний результат все ще відстає від SOTA на 8.1%, що вказує на значний потенціал для поліпшення архітектури моделі та механізмів інтеграції теплової та RGB-інформації. Висока базова точність пов’язана з меншою кількістю класів (4 + background) та більш контрольованими умовами збирання даних у підземних тунелях, що робить сегментацію відносно простішою порівняно з MFNet або SemanticRT.

По-друге, результати для MFNet демонструють mIoU 53.3%, що на 4.6% нижче від SOTA (57.9%, SGFN). Це свідчить про те, що базова модель MFNet не

повністю використовує потенціал RGB-T злиття для різноманітних дорожніх умов, особливо у складних нічних або малоконтрастних сценах. MFNet залишається важливим бенчмарком з 8 класами міських об'єктів для тестування RGB-T методів сегментації.

По-третє, SemanticRT демонструє базовий результат mIoU 79.26%, що на 5.21% нижче від SOTA (84.47%). Незважаючи на те, що це найвищий базовий результат серед розглянутих датасетів, розрив із SOTA вказує на можливості подальшого вдосконалення. Це пояснюється великою різноманітністю умов у датасеті, включно з різними температурними режимами, складним фоновим шумом та високою кількістю класів (12 + background), що ускладнює сегментацію. SemanticRT є найбільшим датасетом (11371 пар зображень – у 7 разів більше, ніж MFNet) з найвищою роздільною здатністю (1280×1024) та найбільшим відсотком анотованих передніх пікселів (21.27%).

Таким чином, моделі для RGB-T сегментації показують суттєві відмінності у точності залежно від датасету та його характеристик. Найвищий базовий результат досягається на SemanticRT (79.26%), що відображає як складність датасету, так і ефективність сучасних методів на великомасштабних даних. PST900 демонструє високу точність (77.5%) на спеціалізованому застосуванні, тоді як MFNet показує найнижчий результат (53.3%) через різноманітність міських сцен та старішу baseline архітектуру. Існує помітний розрив між базовими результатами та SOTA на всіх датасетах (від 4.6% до 8.1%), що підкреслює актуальність удосконалення алгоритмів архітектур та алгоритмів інтеграції спектральних каналів. Найбільш перспективними напрямками є оптимізація моделей для складних різноманітних сцен, підвищення стабільності сегментації у різних умовах освітлення і температури, застосування attention mechanisms та transformer-based архітектур. Важлива також розробка методів, що забезпечують високу точність на великомасштабних різноманітних датасетах для мобільних або автономних платформ.

Порівняння трьох архітектур із використанням результатів, представлених у форматі 2×2, створило можливість детально вивчити їхні сильні й слабкі сторони.

MFNet продемонструвала стабільність у базових сценах, PST900 показала кращу здатність враховувати складні просторові закономірності завдяки глибшій структурі. Навпаки, SemanticRT вирізнялась чіткішими контурами та точнішим окресленням меж об'єктів. Оскільки контури формувалися без додаткової постобробки, ці результати відображали реальну ефективність роботи кожної моделі, дозволяючи робити висновки про їхню придатність до практичного застосування.

Перспективи подальшого розвитку розробленої моделі, орієнтованої на точне відтворення контурів об'єктів у мультимодальних RGB-T зображеннях, охоплюють кілька напрямів, спрямованих на підвищення її структурної точності, універсальності та можливості інтеграції у складніші операційні сценарії. Одним із ключових завдань є покращення узгодження теплових та видимих ознак у приконтурних областях, оскільки саме ці ділянки найчутливіші до шумів, змін освітлення та температурних артефактів [1, 9, 14]. Подальша оптимізація механізмів багаторівневого злиття модальностей дозволить моделі формувати більш стабільні та точні межі навіть у випадках слабкого RGB-сигналу або надмірно розмитого теплового профілю [3, 6]. Іншим перспективним напрямом є розширення набору підтримуваних класів з акцентом на ті категорії, для яких коректне визначення контурів є критично важливим. Додавання таких класів, як вантажні автомобілі, автобуси, тварини, дорожні знаки, перешкоди, будівлі, рослинність чи джерела тепла і диму, дозволить покращити здатність моделі до контурної сегментації у гетерогенних середовищах та збільшить її придатність для реальних сценаріїв моніторингу та відстеження [4, 5].

Важливою складовою подальшої еволюції системи є її інтеграція у практичні застосування, зокрема в безпілотні літальні апарати, для яких якість обводки об'єктів має безпосередній вплив на прийняття рішень у реальному часі. Точне виділення контурів у нічних, закамуфльованих та висококонтрастних сценах покращує навігаційну безпеку дронів. При цьому забезпечується більш надійна оцінка відстані до перешкод, виявлення малоконтрастних чи теплово слабовиражених об'єктів та коректне формування маршрутів в умовах обмеженої

видимості [6]. Інтеграція RGB-T контурної сегментації також підвищує ефективність пошуково-рятувальних операцій, дозволяючи дрону розпізнавати силуети людей та техніки у складних середовищах, де традиційні RGB-дані стають ненадійними – у лісистій місцевості, під час пожеж, у середовищах із димом, туманом чи температурними артефактами. Для високодинамічних польотів модель забезпечує стабільність обрисів навіть за умов значних коливань освітлення, швидких змін траєкторії та впливу атмосферних факторів. Все це зменшує ризик помилкових детекцій та підвищує точність автономних рішень. Завдяки цьому UAV-платформи можуть виконувати моніторинг інфраструктури, патрулювання територій та картографування з підвищеною надійністю, отримуючи структурно точні контури об’єктів у режимі реального часу.

Подальший розвиток моделі передбачає також перехід від обробки окремих кадрів до повноцінної роботи з відеопотоками. Використання темпоральної RGB-T сегментації забезпечить стабільність контурів у динамічних сценах, тоді як впровадження trajectory-aware підходів дозволить відстежувати межі рухомих об’єктів у часі, зберігаючи цілісність їх форми [1, 7]. Це відкриває можливості для задач прогнозування руху, аналізу поведінки та підвищення надійності контурних моделей у сценаріях зі складною динамікою та змінними умовами спостереження [14].

Висновки.

У ході дослідження було проведено комплексний аналіз мультимодального (RGB–T) розпізнавання та класифікації цілей, спрямований на підвищення точності сегментації в умовах змінного або недостатнього освітлення. Особливу увагу приділено принципам злиття спектральних даних від раннього інтегрування ознак до сучасних механізмів середнього та пізнього злиття. Розглянуті також підходи, що застосовують увагу (attention) для вибіркового виділення інформативних компонентів різних модальностей.

Було детально проаналізовано три ключові набори даних: IR Seg Dataset (MFNet), PST900 та SemanticRT. Кожен із них характеризується різними типами сцен, кількістю класів, умовами збирання та рівнем складності. Для забезпечення

коректного порівняння моделі були протестовані на вирівняних і попередньо стандартизованих даних, включно з додатковою обробкою теплових карт, необхідною для IR Seg Dataset. Це забезпечило однакові умови інференсу для всіх розглянутих архітектур.

Результати експериментів продемонстрували суттєву різницю в якості сегментації залежно від архітектурних рішень. MFNet, яка використовує раннє злиття ознак, показала базову стабільність, але обмеженість у складних сценах. PST900 продемонструвала високу точність за рахунок меншої кількості класів та однорідних умов. SemanticRT забезпечила найкращі базові результати завдяки більш глибокому та різноманітному датасету. Загальна продуктивність усіх моделей, однак, виявилася нижчою за сучасні SOTA-підходи, що вказує на наявність простору для оптимізації та подальшого розвитку моделей.

Окремою важливою частиною роботи стало включення та оцінювання контурно-орієнтованої моделі, яка використовує edge supervision для підвищення точності відтворення меж об’єктів. На відміну від класичних моделей, що фокусуються переважно на внутрішній заповненості сегментів, контурно-орієнтований підхід дозволив значно покращити якість локалізації тонких структур, геометричних деталей та складних меж. Це особливо проявилося на складних сценах SemanticRT та на об’єктах із частковим перекриттям або низьким контрастом між класами. Використання додаткового контурного сигналу продемонструвало переваги там, де точність геометрії є критично важливою.

Отримані результати демонструють, що застосування сучасних механізмів злиття модальностей, attention-блоків та контурно-орієнтованих методів є ключовим для підвищення точності мультимодального розпізнавання об’єктів. Аналіз показав, що кожен підхід має свої переваги, а їх поєднання відкриває можливість створювати ще більш ефективні системи. Отримані результати дають розуміння того, які саме архітектурні рішення, механізми злиття даних та методи обробки контурів є найбільш ефективними для побудови таких систем.

У підсумку дослідження формує цілісне бачення того, як має бути

влаштована сучасна система мультимодального розпізнавання на основі штучного інтелекту, та окреслює перспективні напрямки її подальшого вдосконалення. Висновки дозволяють визначити конкретні складові, які потребують покращення. До них належать точніше узгодження RGB і теплових даних, інтеграція attention-модулів та використання контурно-орієнтованих методів. Усе це сприяє розвитку більш ефективних систем класифікації та сегментації цілей у складних умовах спостереження.

Таким чином, наше дослідження робить важливий внесок у розвиток підходів до створення високоточних AI-систем мультимодального розпізнавання та класифікації цілей і створює основу для майбутніх інновацій у цій сфері.

Література

1. Baltrušaitis, T., Ahuja, C., Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. IEEE TPAMI, 2019. DOI: https://doi.org/10.1109/TPAMI.2018.2798607

2. Hazirbas, C., Ma, L., Domokos, C., Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion. ACCV, 2016. DOI: https://doi.org/10.1007/978-3-319-54181-5_14.

3. Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T. MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. IEEE, 2017. DOI: https://doi.org/10.1109/IROS.2017.8206396.

4. Shivakumar, S. S.; Rodrigues, N.; Zhou, A.; Miller, I. D.; Kumar, V.; Taylor, C. J. PST900: RGB-Thermal Calibration, Dataset and Segmentation Network. arXiv, 2019. / GRASP Lab, University of Pennsylvania. arXiv preprint. URL: https://arxiv.org/abs/1909.10980.

5. Ji, W.; Li, J.; Bian, C.; Zhang, Z.; Cheng, L. SemanticRT: Large-Scale RGB-T Dataset for Semantic Segmentation and Target Recognition. arXiv, 2023. DOI: https://doi.org/10.1145/3581783.3611738.

6. Sun, Y.; Zuo, W.; Liu, M. RTFNet: RGB-T Fusion Network for Semantic Segmentation of Urban Scenes. IEEE, 2019. DOI:

https://doi.org/10.1109/LRA.2019.2904733.

7. Chen, Z., et al. MMTM: Multimodal Transfer Module for CNN Fusion. CVPR, 2020. DOI: https://doi.org/10.48550/arXiv.1911.08670.

8. Li, P.; Chen, J.; Lin, B.; Xu, X. Residual Spatial Fusion Network for RGB-Thermal Semantic Segmentation. In: Proceedings / preprint, 2023. URL: https://huggingface.co/papers/trending.

9. Li, G.; … CAFNet: Cross-Modal Adaptive Fusion Network With Attention and Gated Weighting for RGB-T Semantic Segmentation. IEEE Access, (DOAJ). DOI: https://doi.org/10.1109/access.2025.3595811.

10. Zhou, Z.; Wu, S.; Zhu, G.; Wang, H.; He, Z. Channel and Spatial Relation-Propagation Network for RGB-Thermal Semantic Segmentation (CSRPNet). arXiv, 2023. DOI: https://doi.org/10.48550/arXiv.2308.12534.

11. He, K., Girshick, R., Dollár, P. Feature Pyramid Networks for Object Detection. CVPR, 2017. DOI: https://doi.org/10.1109/CVPR.2017.106.

12. Valada, A., Vertens, J., Dhall, A., Burgard, W. Self-Supervised Multimodal Fusion for Semantic Segmentation. IEEE RAL, 2019. DOI: https://doi.org/10.1007/s11263-019-01188-y.

13. Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; Stiefelhagen, R. CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation with Transformers. arXiv, 2022. DOI: https://doi.org/10.48550/arXiv.2203.04838.

14. Wang, M.; Zhu, Z.; Wang, Y.; Tu, R.; Weng, J.; Yu, X. Edge-Supervised Attention-Aware Fusion Network for RGB-T Semantic Segmentation. Electronics, 2025. DOI: https://doi.org/10.3390/electronics14081489.

Abstract. *The study addresses the problem of developing a multimodal system for target recognition and classification based on artificial intelligence. The relevance of the topic stems from the need to improve the reliability of automatic object detection under conditions of low visibility, varying illumination, interference, and diverse atmospheric environments. To achieve this, multimodal approaches combining information from the visible (RGB) and thermal (IR/TIR) spectra are employed, which significantly increase the robustness and accuracy of the models.*

The research analyzes three representative datasets–MFNet/ir_seg, PST900 RGB-T, and SemanticRT–which provide aligned RGB and thermal images for tasks of segmentation, detection, and object classification. A review of current AI-based multimodal deep learning approaches is conducted, including methods of early, mid, and late fusion, as well as models with edge-oriented supervision. Based on the analysis, the study identifies which architectures are most effective for

target classification under different conditions–from daytime and nighttime observation scenarios to low-light environments where thermal data becomes particularly important.

The results of the work provide a systematic understanding of the capabilities of artificial intelligence in multimodal recognition and help determine the optimal architectures and models for further experimental research.

Key words: *multimodal recognition, target classification, artificial intelligence, RGB-T data, deep learning, edge-aware learning, semantic segmentation.*