

DOI: <https://doi.org/10.15276/ict.03.2026.55>

УДК 004.932:004.85

Сегментація та детекція людей у мультиспектральному відеопотоці в умовах надзвичайних ситуацій

Угрин Дмитро Ілліч¹⁾

ORCID: <https://orcid.org/0000-0003-4858-4511>; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300

Бреславський Олег Ігорович¹⁾

ORCID: <https://orcid.org/0009-0005-5164-2913>; breslavskyi.oleh@chnu.edu.ua

¹⁾ Чернівецький національний університет імені Юрія Федьковича, вул. Коцюбинського, 2.
Чернівці, 58012, Україна

АНОТАЦІЯ

Роботу присвячено підвищенню точності та стійкості детекції й сегментації людей у мультиспектральному відеопотоці за умов надзвичайних ситуацій зі збереженням продуктивності, достатньої для роботи в реальному часі. Показано, що пряме перенесення детекторів, навчених на зображеннях видимого діапазону, в інфрачервоний спектр є неефективним через міжмодальний доменний зсув: модальності різняться за фізичною природою сигналу, рівнем шуму, характером контрасту та набором доступних текстурних ознак. Проаналізовано основні напрями детекції та сегментації людей в інфрачервоному діапазоні: теплові детектори та спеціалізовані набори даних, доменну адаптацію й перенесення ознак із видимого спектра, багатоспектральні детектори з різними рівнями злиття та методи компенсації модального дисбалансу. Встановлено, що їх спільним обмеженням є фіксована довіра до модальностей та однакова для всіх кадрів глибина попередньої обробки. Запропоновано вдоскоалений спосіб детекції та сегментації людей на основі двогілкової архітектури YOLOv8m-seg, що поєднує кількісну метрику інформативності модальності, адаптивну передобробку кадрів і крос-модальне злиття ознак зі змінними вагами. Інформативність RGB- та ІЧ-каналів визначається за нормованими характеристиками локального середньоквадратичного контрасту, ентропії розподілу яскравості, енергії градієнта за Тененграмом, оберненої оцінки цільності серпанку та відношення сигнал/шум, а для теплового каналу додатково враховується узгодженість теплового діапазону цілі. Значення метрики належать діапазону $[0; 1]$ і поділяються на чотири рівні: дуже низький за $I < 0,25$, низький за $0,25 \leq I < 0,45$, середній за $0,45 \leq I < 0,70$ та високий за $I \geq 0,70$. За $I(c) < 0,45$ режим втручання визначається профілем компонент метрики: залежно від характеру деградації застосовується швидка корекція або нейромережеве відновлення, тоді як за достатньої інформативності додаткова передобробка пропускається. Крос-модальне злиття реалізовано зі змінними просторовими вагами RGB- та ІЧ-ознак, що визначаються відповідно до локальної інформативності кожної модальності. Додатково застосовано узгоджену з фізикою деградацій стратегію формування навчальної вибірки та засоби роботи з малорозмірними й частково перекритими цілями. Експериментальну перевірку проведено на відкритих мультиспектральних наборах KAIST Multispectral Pedestrian, LLVIP та FMB, об'єднаних у відтворюваний протокол MS-ER; матеріали для відтворення експерименту розміщено у відкритому репозиторії GitHub. Експериментальна база охоплювала 26 841 пару RGB-ІЧ кадрів, або 53 682 зображення, а для навчання сегментаційної складової додатково сформовано 55 427 масок людей. Для контрольованого порівняння використано шість конфігурацій M1–M6: RGB-only, thermal-only, статичне RGB+ІЧ-злиття, ICAFusion, запропоновану модель без адаптивної передобробки та повну модель M6 з метрикою інформативності, адаптивною передобробкою й локальним керуванням вагами модальностей. За результатами експериментального прогону 5000 кадрів встановлено, що 46,8 % кадрів оброблялися без додаткової передобробки, для 31,6 % застосовувалася швидка корекція, а для 21,6 % – глибока реконструкція; середній час обробки становив 26,3 мс/кадр за середньої продуктивності 38,0 FPS, максимальної 49,8 FPS та пікового використання GPU-пам'яті 6,4 ГБ. Отримані результати підтверджують працездатність запропонованого способу та можливість адаптивного перерозподілу обчислювальних ресурсів залежно від поточної інформативності RGB- та ІЧ-каналів. Застосування найбільш ресурсомісткого режиму лише до деградованої частини кадрів дає змогу зберігати продуктивність, достатню для обробки мультиспектрального відеопотоку в реальному часі, водночас використовуючи комплементарну інформацію RGB- та ІЧ-модальностей для детекції й сегментації людей у складних умовах спостереження.

Ключові слова: мультиспектральний відеопотік; детекція людей; семантична сегментація; інфрачервоний діапазон; міжмодальний доменний зсув; крос-модальне злиття ознак; метрика інформативності модальності; адаптивна передобробка; малорозмірні цілі; обробка в реальному часі

Для цитування: Угрин Д. І., Бреславський О. І. “Сегментація та детекція людей у мультиспектральному відеопотоці в умовах надзвичайних ситуацій”. *Інформатика. Культура. Техніка*. 2026; Том 3 № 2 (4): 718–733. DOI: <https://doi.org/10.15276/ict.03.2026.55>

Актуальність. Сучасні системи ідентифікації людей у відеопотоці демонструють високу точність на зображеннях видимого діапазону, отриманих за нормального освітлення. Проте в умовах надзвичайних ситуацій – пожеж, задимлення, туману, нічних пошуково-рятувальних операцій – інформативність RGB-каналу різко падає, і продуктивність таких моделей деградує. Природним рішенням є залучення інфрачервоного (ІЧ) спектра, передусім довгохвильового теплового діапазону, у якому людина реєструється за власним

випромінюванням і не залежить від зовнішнього освітлення. Однак пряме перенесення моделі, навченої на RGB-даних, в ІЧ-домен не дає очікуваного результату [1], [2].

Причиною є міжмодальний доменний зсув: розподіли ознак двох модальностей суттєво відрізняються. ІЧ-зображення не містять кольору й більшості текстур, мають інший рівень шуму, інше співвідношення сигнал/шум та інший характер контрасту, зумовлений тепловою сигнатурою, а не відбитим світлом. Низькорівневі фільтри, навчені реагувати на кольорові переходи та дрібну текстуру одягу, у тепловому кадрі активуються на випадкових перепадах температури фону, тому точність детектора падає навіть тоді, коли ціль добре помітна для оператора [1], [2], [3].

Додаткову складність створює специфіка самих цілей. У кадрах, отриманих з безпілотних носіїв або стаціонарних камер спостереження, людина на значній відстані займає лише десятки пікселів, що робить її малорозмірною ціллю з мінімальним набором ознак. Часткові оклюзії уламками, конструкціями, іншими людьми чи щільними клубами диму руйнують цілісність силуету. Нерівномірний тепловий фон під час пожежі маскує теплову сигнатуру людини гарячими поверхнями, а рух носія додає змазування. Сукупність цих факторів призводить до пропуску цілей і хибних спрацювань саме в тих сценаріях, де ціна помилки максимальна [4], [5].

Тому актуальним є розроблення способу детекції та сегментації людей, стійкого до зазначених деградацій, здатного оцінювати придатність кожної модальності в поточних умовах і придатного до роботи в режимі реального часу.

Мета дослідження. Метою дослідження є розроблення та експериментальна перевірка способу детекції й сегментації людей у мультиспектральному RGB–ІЧ відеопотоці в умовах надзвичайних ситуацій на основі адаптивного оцінювання інформативності модальностей, керованої передобробки та крос-модального злиття ознак для підвищення точності й стійкості виявлення людей за низької видимості, задимлення, туману, теплових завад, а також для малорозмірних і частково перекритих цілей зі збереженням продуктивності, достатньої для роботи в реальному часі. Для досягнення поставленої мети сформульовано такі завдання: розробити кількісну метрику оцінювання глобальної та локальної інформативності RGB- та ІЧ-модальностей; реалізувати адаптивну передобробку з автоматичним вибором між пропуском обробки, швидкою корекцією та нейромережевим відновленням залежно від характеру деградації; розробити двогілкову архітектуру детектора-сегментатора на основі YOLOv8m-seg із крос-модальним злиттям та просторово змінними вагами RGB- й ІЧ-ознак; сформувати стратегію навчання, орієнтовану на модально-специфічні деградації, малорозмірні та частково перекриті цілі; реалізувати темпоральну обробку з виділенням ключових кадрів і використанням трекінгу для зменшення обчислювальних витрат; провести експериментальну перевірку на відкритих мультиспектральних наборах KAIST, LLVIP та FMB; оцінити точність детекції й сегментації, стійкість до різних типів деградацій та обчислювальну продуктивність запропонованого способу й визначити його придатність до обробки відеопотоку в реальному часі.

Аналіз сучасних досліджень. Дослідження детекції та сегментації людей в інфрачервоному діапазоні умовно поділяються на чотири напрями.

Перший напрям – суто теплові детектори, навчені безпосередньо на ІЧ-даних. Їхній розвиток тривалий час стримувався браком розмічених наборів; ситуацію змінила поява спеціалізованих датасетів із попільсько вирівняними RGB/ІЧ-парами, зокрема KAIST для пішохідної детекції в міському середовищі [4] та LLVIP, орієнтованого на сцени з дуже низьким рівнем освітлення [6]. Ці набори зробили можливим коректне порівняння модальностей і показали, що тепловий канал сам по собі забезпечує вищу повноту вночі, проте помітно програє видимому за наявності теплових завад і в задачах розрізнення близько розташованих цілей, силуети яких зливаються в спільну теплову пляму [6].

Другий напрям – доменна адаптація та перенесення ознак із видимого спектра в тепловий. Оскільки обсяг розмічених ІЧ-даних залишається на порядки меншим за обсяг RGB-наборів, поширеною стратегією є використання RGB-моделі як джерела апіорних

знань із подальшим донавчанням на теплових кадрах. Досліджено підходи, у яких із видимого спектра запозичуються псевдомодальні представлення для компенсації браку текстур у тепловому зображенні [3], а також схеми адаптації, у яких вирівнювання доменів виконується з урахуванням цільової задачі детекції, а не лише статистики ознак [5]. Ці роботи підтверджують наявність доменного зсуву як самостійної проблеми та показують, що його неможливо усунути простим донавчанням останніх шарів мережі.

Третій напрям – багатоспектральні детектори, що використовують обидві модальності одночасно. Тут сформувалася усталена класифікація за рівнем злиття: раннє злиття на рівні вхідного тензора, злиття на середньому рівні ознак і пізнє злиття вже готових прогнозів двох незалежних гілок [7]. Порівняльні дослідження показують перевагу злиття на середньому рівні як компромісу між глибиною взаємодії модальностей і обчислювальною вартістю [7], а подальший розвиток напряму пов'язаний із введенням механізмів уваги, які дозволяють мережі перерозподіляти внесок гілок замість простої конкатенації ознак [8]. Близькими до цієї ідеї є підходи з урахуванням освітленості сцени, у яких окрема підмережа оцінює ймовірність денних або нічних умов і відповідно до неї зважує прогнози RGB- та ІЧ-гілок [9], [10]. Проте в таких рішеннях ваги визначаються єдиною глобальною оцінкою освітленості кадру, тоді як інші типи деградацій (дим, туман, теплові завади) та їх просторова неоднорідність у межах кадру не враховуються.

Четвертий напрям – компенсація модального дисбалансу. Установлено, що навчена на змішаних даних мережа схильна систематично надавати перевагу тій модальності, яка домінує в навчальній вибірці, унаслідок чого друга гілка недонавчається й перестає робити внесок у прогноз навіть тоді, коли є єдиним джерелом корисного сигналу [11]. Запропоновані засоби протидії – роздільне контролювання гілок та регуляризація їхнього внеску – усувають дисбаланс на рівні навчання, проте залишають ваги модальностей сталими під час висновування.

Окремо слід відзначити роботи з відновлення зображень за умов низької освітленості та атмосферних завад [12], [13], які традиційно розглядаються поза контекстом детекції й застосовуються до всієї послідовності однаково, без урахування того, чи потребує конкретний кадр такої обробки взагалі.

Отже, спільним обмеженням розглянутих підходів є те, що довіра до модальності або фіксується на етапі проектування системи, або визначається лише глобальною освітленістю кадру, а глибина попередньої обробки залишається однаковою для всіх кадрів. Придатність RGB- та ІЧ-каналів у реальних сценаріях надзвичайних ситуацій змінюється в межах однієї послідовності – від нічної сцени, де домінує тепловий канал, до кадрів із гарячим фоном пожежі, де він деградує першим. Це формує невирішену раніше частину проблеми: потрібен механізм, який кількісно оцінює поточну інформативність кожної модальності та керує вагами злиття за інтегральною оцінкою, а обсягом передобробки – за профілем її компонент.

Виклад основного матеріалу дослідження. Пропонована робота будується навколо однієї центральної ідеї: довіра до модальності має бути вимірюваною величиною, обчислюваною для кожного кадру, а не сталою, зафіксованою під час навчання. Із цієї ідеї послідовно випливають усі три складники запропонованого способу – метрика інформативності, адаптивна передобробка та злиття ознак зі змінними вагами.

Для експериментальної перевірки запропонованого способу сформовано уніфікований протокол порівняння шести конфігурацій детекції та сегментації людей:

M_1 – RGB-only; M_2 – thermal-only; M_3 – RGB+ІЧ зі статичним середньорівневим злиттям ознак; M_4 – сучасне крос-модальне злиття ICAFusion [8]; M_5 – запропонована двогілкова модель із крос-модальним злиттям, але без адаптивної передобробки; M_6 – повна запропонована модель із метрикою інформативності $I(c)$, адаптивною передобробкою та локальним керуванням вагами RGB- та ІЧ-ознак за картою $I(c, x, y)$.

Конфігурації M_1 і M_2 використовували для визначення внеску кожної модальності окремо, M_3 – як контроль статичного мультиспектрального злиття; M_4 – як зовнішній сучасний baseline; M_5 – для абляційного визначення внеску адаптивної передобробки, а M_6 – як повну

реалізацію запропонованого способу. Вище наведена постановка безпосередньо відповідає структурі розробленої архітектури, у якій передбачено окремі RGB- та ІЧ-енкодера, середньорівневе злиття та керування внеском модальностей за їх локальною інформативністю.

Експерименти проводили на піпксельно узгоджених парах RGB/ІЧ-зображень. Вихідний набір даних розділяли на навчальну, валідаційну та тестову частини у співвідношенні 70:15:15 на рівні відеопослідовностей, а не окремих кадрів, щоб кадри однієї сцени не потрапляли одночасно до навчальної та тестової вибірок. Для всіх шести конфігурацій використовували той самий поділ даних. RGB- та ІЧ-кадри однієї пари піддавали однаковим геометричним перетворенням, щоб не порушувати їх просторове узгодження. Модально-специфічні фотометричні деградації застосовували лише до відповідного спектрального каналу. Такий принцип відповідає прийнятій у роботі стратегії, за якої дим, туман, шум, змазування, зниження роздільної здатності та теплові завади моделюються з урахуванням фізики відповідної модальності.

Для виключення впливу різної обчислювальної потужності моделей у конфігураціях M_1 – M_3 , M_5 і M_6 використовували однаковий базовий backbone (використовували однакову базову архітектуру енкодера) та однакові голови детекції й сегментації

Для M_1 активною залишали лише RGB-гілку, для M_2 – лише ІЧ-гілку, у M_3 ознаки двох гілок об'єднували на тому самому рівні, що й у запропонованій моделі, але з фіксованими вагами $w_{RGB}=w_{IR}=0,5$. У M_5 зберігали запропонований механізм адаптивного крос-модального злиття, проте модуль адаптивної передобробки повністю вимикали. У M_6 використовували всі компоненти запропонованого способу: розрахунок глобального $I(c)$, локальної карти $I(c,x,y)$, вибір режиму передобробки та адаптивне злиття. Для M_4 реалізовували ICAFusion відповідно до оригінальної архітектури [8], яка вже включена до теоретичної бази роботи як сучасний метод крос-модального злиття.

Для M_6 рішення про передобробку приймали до виконання основного прямого проходу. За $I(c) \geq 0,45$ канал передавали до енкодера без додаткової передобробки, а за $I(c) < 0,45$ тип втручання визначали за профілем компонент метрики. За переважання ознак розсіювання або зниження відношення сигнал/шум застосовували нейромережеву реконструкцію, а за тональним погіршенням або інверсією теплового контрасту – швидку корекцію. Таким чином, поріг та логіку перемикання режимів у процесі тестування не змінювали й не оптимізували за тестовими даними. Значення $I < 0,25$, $0,25 \leq I < 0,45$, $0,45 \leq I < 0,70$, та $I \geq 0,70$ відповідали дуже низькому, низькому, середньому та високому рівням інформативності відповідно, як визначено в запропонованому способі.

Тестову вибірку оцінювали, як у цілому, так і окремо для семи підмножин: нормальні умови; ніч/слабке освітлення; дим; туман; гарячий фон; одночасний гарячий фон і задимлення; складні цілі. Останню категорію додатково поділяли на малорозмірні та частково перекриті цілі. Малорозмірною вважали ціль із площею обмежувального прямокутника менше 32^2 пікселів після приведення кадру до єдиного вхідного розміру, що відповідає визначенню малих об'єктів у протоколі COCO [14]. До перекритих відносили ановані цілі, для яких частина силуету закрита іншим об'єктом; для забезпечення відтворюваності додатково виділяли підмножину з перекриттям не менше 30% площі силуету. Такий окремий аналіз необхідний, оскільки саме малий розмір і часткове перекриття в роботі визначено як фактори, що спричиняють втрату характерних ознак людини.

Для задачі детекції для кожної конфігурації визначали Precision, Recall, F1, mAP@0.5 та mAP@0.5:0.95. Значення TP і FP встановлювали за відповідністю прогнозованого та еталонного bounding box при $IoU \geq 0,50$; для mAP@0.5:0.95 використовували пороги IoU від 0,50 до 0,95 із кроком 0,05 відповідно до протоколу COCO [14]. Для сегментації визначали IoU кожного прогнозованого силуету та середній mIoU тестової вибірки. Для малих цілей окремо розраховували AP_s , $Recall_s$ і $F1_s$, а для частково перекритих – AP_{occ} , $Recall_{occ}$, $F1_{occ}$ та $mIoU_{occ}$. Таким чином, покращення інтегрального mAP не розглядали, як достатній доказ

ефективності методу без одночасної перевірки його поведінки на малорозмірних та перекритих цілях, для роботи з якими у запропонованій схемі передбачено спеціальне формування навчальної вибірки та використання сегментаційної гілки.

Кожну конфігурацію M_1 – M_6 навчали п'ять разів із незалежними початковими ініціалізаціями $seed=\{11,23,37,51,79\}$, зберігаючи незмінними розбиття даних та всі гіперпараметри. Підсумковий результат для кожної метрики визначали, як середнє значення п'яти запусків із наведенням стандартного відхилення $\bar{x} \pm SD$. Для $mAP@0.5:0.95$, $mIoU$, AP_s та AP_{occ} додатково обчислювали 95% довірчі інтервали методом непараметричного bootstrap із 2000 повторних вибірок тестових відеопослідовностей. Повторну вибірку виконували саме на рівні послідовностей, а не окремих сусідніх кадрів, щоб зменшити вплив часової кореляції відеоданих. Для порівняння M_6 із M_3 , M_4 і M_5 розраховували парні різниці ΔmAP , $\Delta mIoU$, ΔAP_s та ΔAP_{occ} на однакових тестових послідовностях і відповідні 95% bootstrap-довірчі інтервали. Перевагу вважали статистично стійкою, якщо довірчий інтервал відповідної різниці не містив нульового значення.

Оцінювання швидкодії виконували на одному й тому самому апаратному забезпеченні за $batch\ size = 1$, що відповідає послідовній обробці відеопотоку. Після 100 прогрівальних проходів виконували не менше 1000 вимірювань без урахування часу завантаження даних із накопичувача. Для кожної конфігурації визначали середній час обробки кадру, 95-й перцентиль затримки та FPS. Для M_6 окремо реєстрували час розрахунку $I(c)$, час передобробки, час прямого проходу детектора-сегментатора та крос-модального злиття, а також частку кадрів, для яких передобробку було пропущено, застосовано швидку корекцію, або нейромережеву реконструкцію. Це дозволяє експериментально перевірити сформульоване в роботі положення про те, що середня обчислювальна вартість визначається не максимальною складністю реконструкції, а частотою її активації.

У темпоральному режимі додатково фіксували загальну кількість кадрів відеопослідовності N і кількість ключових кадрів K , для яких виконувався повний прохід конвеєра, та визначали частку ключових кадрів як K/N . Ключовий кадр призначали при перевищенні встановленого порогу зміни $I(c)$ відносно попереднього ключового кадру, що також відповідає вже описаному алгоритму.

В результаті сформований протокол забезпечував три рівні контрольованого порівняння: M_1 – M_2 проти M_3 – M_6 визначає ефект використання комплементарної інформації RGB та ІЧ; M_3 – M_4 проти M_6 визначає ефект адаптивного керування внеском модальностей порівняно зі статичним і сучасним крос-модальним злиттям; M_5 проти M_6 ізолює внесок саме адаптивної передобробки. При цьому, даний протокол забезпечує контрольоване порівняння запропонованого способу з RGB-only, thermal-only, статичним RGB+ІЧ-злиттям та сучасним крос-модальним методом, а абляція M_5 – M_6 ізолює внесок адаптивної передобробки.

Відмітимо, що у відповідності до [1], [3] загальні mAP , Precision, Recall, F1 і $mIoU$ характеризують інтегральну ефективність, AP_s та AP_{occ} – стійкість до малорозмірності й оклюзії, стратифіковане тестування – стійкість до конкретних типів деградації, а 95 % bootstrap-довірчі інтервали – статистичну відтворюваність отриманих відмінностей. При цьому окреме оцінювання AP_s , AP_{occ} і $mIoU_{occ}$, стратифікація за типами деградації та 95 % довірчі інтервали забезпечують перевірку не лише середньої точності, а й статистичної та сценарної стійкості запропонованого способу.

На методичному рівні суть запропонованого способу базується на тому що для модальності s формується інтегральний показник $I(c)$ як зважена сума нормованих компонент: локального середньоквадратичного контрасту, ентропії розподілу яскравості, енергії градієнта за оператором Тененграма, оберненої оцінки щільності серпанку за методом темного каналу [15], оцінки відношення сигнал/шум за різницею кадру та його медіанно згладженої версії та узгодженості теплового діапазону цілі.

Кожна компонента приводиться до діапазону $[0; 1]$ за статистикою, зібраною на калібрувальній вибірці, після чого показник обчислюється за виразом

$$I(c) = \sum_k w_k \cdot f_k(c), \quad \sum_k w_k = 1, \quad I(c) \in [0; 1], \quad (1)$$

де $f_k(c)$ – k -та нормована компонента, w_k – її вага. Ваги визначаються один раз на калібрувальній вибірці шляхом максимізації кореляції між $I(c)$ та повнотою детекції, отриманою на цій самій модальності. У разі зміни базового детектора або сенсорного комплексу перекалібрування виконується автоматично на тій самій калібрувальній вибірці й не пов'язане зі зміною архітектури.

Для теплового каналу вводиться шоста компонента $f_6(ІЧ)$ – частка пікселів кадру, яскравість яких потрапляє в очікуваний діапазон теплової сигнатури людини $[T_{\min}; T_{\max}]$, нормована на очікувану частку присутності цілі в сцені. Діапазон визначається один раз на калібрувальній вибірці як інтервал, у який потрапляє більшість пікселів анотованих силуетів людей. Зі зростанням частки «надгарячих» пікселів (полум'я, розігріті поверхні) або «надхолодних» (охолоджені конструкції взимку) значення $f_6(ІЧ)$ падає, оскільки корисний сигнал витісняється з робочого діапазону. Компонента f_6 застосовна лише до теплового каналу як спосіб урахування фізики теплової сигнатури; для видимого каналу показник $I(RGB)$ обчислюється за п'ятьма фотометричними компонентами, ваги яких нормуються до одиничної суми окремо.

Показник (1) обчислюється у двох режимах: глобальному, коли всі п'ять фотометричних компонент і компонента f_6 усереднюються по всьому кадру та дають скалярне значення $I(c)$, і локальному, коли ті самі компоненти обчислюються на регулярній сітці блоків розміром $B \times B$ і дають карту $I(c, x, y)$, у якій кожне значення відповідає інформативності модальності c у ділянці з координатами (x, y) . Скалярний $I(c)$ збігається із середнім значенням локальної карти $I(c, x, y)$, що гарантує узгодженість між двома контурами керування: глобальним, який задає режим передобробки для каналу в цілому, і локальним, який задає ваги злиття для кожної просторової ділянки.

Поділ на якісні рівні, що використовуються далі, задається чотирма порогоми:

- $I < 0,25$ – дуже низька інформативність;
- $0,25 \leq I < 0,45$ – низька;
- $0,45 \leq I < 0,70$ – середня;
- $I \geq 0,70$ – висока.

Такий поділ усуває суб'єктивність оцінок і робить правила вибору режиму передобробки відтворюваними та придатними до автоматичного перенастроювання під конкретний сенсорний комплекс. Числові значення порогів обрано за точкою помітного спаду повноти детекції на калібрувальній вибірці; додатково перевірено, що вони приблизно відповідають квартилям розподілу $I(c)$ і слабо залежать від сенсорного комплексу, а отже, придатні до перенесення на близькі за характеристиками сцени без повторного підбору.

Для визначення параметрів метрики інформативності було сформовано окрему калібрувальну вибірку обсягом 2400 синхронізованих RGB–ІЧ пар кадрів, які не включали до тестової вибірки. Вона охоплювала нормальні умови спостереження, нічне та слабе освітлення, задимлення, туман, гарячий фон і комбіновані деградації. За результатами калібрування для RGB-каналу отримано такі вагові коефіцієнти: локальний середньоквадратичний контраст: $w_1=0,24$; ентропія яскравості: $w_2=0,18$; енергія градієнта за Тененграмом: $w_3=0,23$; обернена оцінка щільності серпанку: $w_4=0,20$; відношення сигнал/шум: $w_5=0,15$. Для ІЧ-каналу відповідні коефіцієнти становили $w_1=0,18$, $w_2=0,13$, $w_3=0,18$, $w_4=0,10$, $w_5=0,16$, а для компоненти узгодженості теплового діапазону цілі – $w_6=0,25$. Вагові коефіцієнти визначали на калібрувальній вибірці за критерієм максимального зв'язку інтегральної інформативності з повнотою детекції відповідної модальності та після калібрування фіксували без подальшого коригування за тестовими даними.

Основою запропонованого способу є двогілкова архітектура з окремими енкодерами для RGB- та ІЧ-потоків і злиттям на середньому рівні ознак (за базову архітектуру детектора-сегментатора прийнято двогілкову модель на основі YOLOv8m-seg [16], модифіковану для

паралельного опрацювання синхронізованих RGB- та ІЧ-кадрів). Вхідні зображення обох модальностей приводили до розміру 640×640 пікселів. RGB- та ІЧ-потoki оброблялися окремими енкодерами однакової структури без спільних ваг. Такий підхід забезпечував незалежне формування ознак для модальностей із різною фізичною природою сигналу, що відповідає визначеному в роботі міжмодальному доменному зсуву. Злиття виконували на середньому рівні представлення ознак, як передбачено вихідною структурою запропонованого способу. Для кожної модальності енкодер формував багатомасштабні карти ознак $FRGB(l)$ та $FICh(l)$ на трьох рівнях l зі stride 8, 16 і 32, що для входу 640×640 відповідало просторовим розмірам 80×80 , 40×40 і 20×20 . Перед злиттям кількість каналів парних карт узгоджували згорткою 1×1 .

Локальну карту інформативності $I(c, x, y)$ формували на регулярній сітці неперекривних блоків $B \times B$, прийнявши $B=32$ пікселів. У кожному блоці незалежно розраховували ті самі нормовані компоненти $fk(c)$, що й для глобального показника $I(c)$, після чого у відповідності до [3] визначали $I(c, x, y)$.

Після злиття багатомасштабні карти передавалися до пещі і спільних вихідних гілок. Детекційна голова формувала координати обмежувальних прямокутників і ймовірність класу «людина», а сегментаційна – попиксельну маску цілі. Спільне використання детекції та сегментації особливо важливе для близько розташованих і частково перекритих людей, оскільки в ІЧ-зображенні їх теплові області можуть зливатися.

Ключова відмінність від розглянутих вище рішень полягає у відмові від статичних ваг модальностей: внесок кожного потоку визначається модулем крос-модальної уваги окремо для кожної просторової ділянки та каналу ознак, а вхідним сигналом для цього модуля є локальна карта інформативності $I(c, x, y)$. Тобто вага модальності є функцією її локальної інформативності: у зоні щільного диму домінує ІЧ-гілка, у затінених ділянках зі збереженою текстурою – RGB-гілка. Спільна голова формує як обмежувальні прямокутники, так і полегшену маску сегментації, що дозволяє відокремити силует людини від фону навіть за розмитих меж об'єкта та розділити цілі, теплові сигнатури яких зливаються. Узагальнену структуру наведено на Рис. 1.

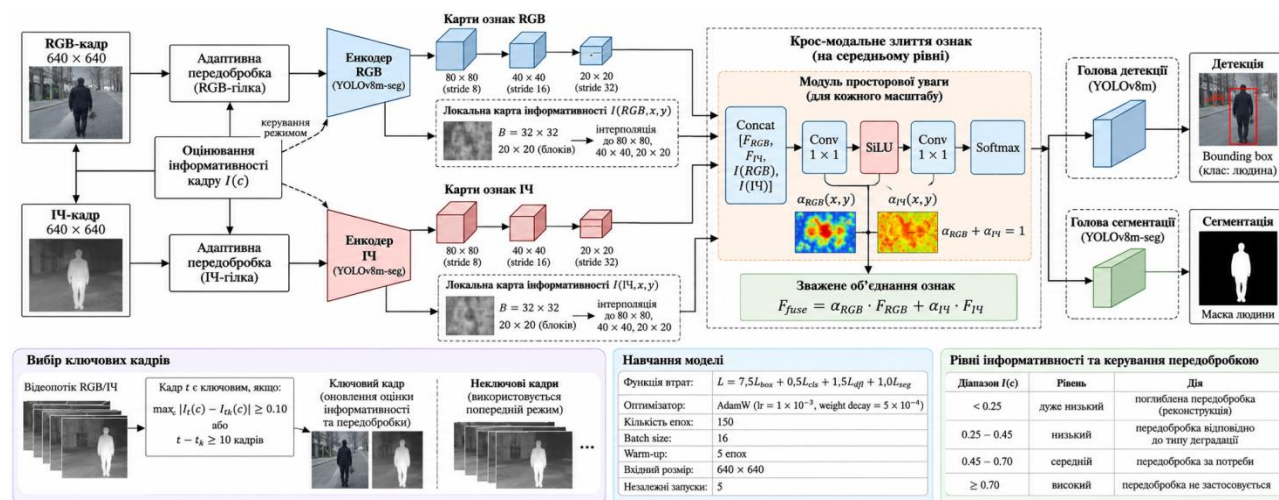


Рис. 1. Двогілкова архітектура з адаптивною передобробкою та крос-модальним злиттям ознак
Джерело: розроблено авторами

При цьому наведена на Рис.1. архітектура реалізує паралельну обробку RGB- та ІЧ-кадрів окремими енкодерами YOLOv8m-seg із подальшим крос-модальним злиттям ознак. Її ключовою особливістю є адаптивне визначення інформативності кожної модальності та формування ваг α_{RGB} і α_{ICh} , завдяки чому внесок RGB та ІЧ даних змінюється залежно від умов сцени. Додатково передбачено керовану попередню обробку та вибір ключових кадрів для зменшення обчислювальних витрат. Об'єднані ознаки використовуються одночасно для детекції людей і їх сегментації.

Для адаптивного злиття використовували модуль крос-модальної уваги. На кожному рівні ознаки RGB- та ІЧ-гілок разом із відповідними значеннями локальної інформативності подавали на послідовність $\text{Conv } 1 \times 1 \rightarrow \text{SiLU} \rightarrow \text{Conv } 1 \times 1 \rightarrow \text{Softmax}$. На виході формували просторово залежні коефіцієнти $\alpha_{\text{RGB}}(x,y)$ та $\alpha_{\text{ІЧ}}(x,y)$, сума яких дорівнювала одиниці.

Внесок кожної модальності змінювався залежно від стану конкретної ділянки кадру, а не задавався фіксованим коефіцієнтом. Сам принцип формування змінних ваг модальностей на основі локальної інформативності передбачений вихідною методикою.

У видимому спектрі маска потрібна переважно для уточнення меж об'єкта, тоді як у тепловому вона виконує додаткову функцію – розділення цілей. Через відсутність кольору й текстури дві людини, що стоять поруч, у тепловому кадрі утворюють єдину зв'язну область із подібною яскравістю, і детектор, який працює лише з обмежувальними прямокутниками, схильний об'єднати їх в одне виявлення. Попіксельна маска, навчена спільно з гілкою детекції, використовує тонкі перепади температури на межі силуетів і дозволяє розділити такі області. З тієї самої причини маска є стійкішою за прямокутник за часткової оклюзії: видима частина силуету зберігає форму, тоді як описаний прямокутник спотворюється перекриттям. Тому в запропонованому способі сегментація розглядається не як додаткова функція, а як засіб компенсації браку ознак у тепловому каналі.

При формуванні навчальної вибірки за базис приймаються публічні мультиспектральні набори з попиксельно вирівняними RGB/ІЧ-парами, які доповнюються керованою аугментацією деградацій: синтезом диму й туману за фізичною моделлю атмосферного розсіювання, адитивним і мультиплікативним шумом, змазуванням від руху, зниженням роздільної здатності та симуляцією теплових завад. Принципово важливим є те, що деградація застосовується модально узгоджено: дим і туман суттєво спотворюють RGB-канал і значно слабше впливають на тепловий, натомість гарячий фон уражає саме тепловий канал, не торкаючись видимого. Це привчає модель довіряти тій гілці, яка в поточних умовах справді інформативна. Для складних цілей додатково застосовується перевибірка малорозмірних об'єктів із перенесенням їхніх екземплярів на нові фони зі збереженням масштабу, а також синтез часткових оклюзій, що виконується по силуету об'єкта, а не по описаному прямокутнику. Це формує вибірку, збалансовану за розміром цілі та ступенем її перекриття.

Стійкість до шумів і деградацій забезпечується на етапі навчання. Модально-специфічний dropout – випадкове вимкнення однієї з гілок – не дозволяє моделі колапсувати при відмові чи засліпленні одного сенсора й водночас протидіє модальному дисбалансу, змушуючи кожен гілку формувати самодостатнє представлення. Узгоджувальна функція втрат вимагає близькості прогнозів на чистому та штучно деградованому кадрі, що робить представлення ознак інваріантним до типу спотворення. Для компенсації дисбалансу між великими й малорозмірними цілями застосовується фокальна функція втрат [17].

Режим реального часу досягається за рахунок адаптивної попередньої обробки. Легкий блок оцінювання стану кадру обчислює для кожної модальності вектор нормованих компонент $f_k(c)$ та їхню згортку $I(c)$ за виразом (1). Скалярний показник визначає, чи потребує канал втручання взагалі, а режим обробки обирається за профілем компонент: він є функцією вектора $f_k(c)$, а не самого лише значення $I(c)$. Такий поділ ролей не потребує додаткових обчислень: усі компоненти вже отримано за один прохід по зменшеній копії кадру, причому оцінка щільності серпанку відокремлює розсіювальну природу деградації від тональної, а компонента узгодженості теплового діапазону – від інверсії теплового контрасту. Оброблюваними вважаються канали з глобальним $I(c) < 0,45$; для кожного з них обирається один із трьох режимів: пропуск передобробки, швидка корекція (вирівнювання гістограми, гамма-корекція), або нейромережеве відновлення. Останнє викликається лише за перевищення порогу деградації, що вивільняє часовий бюджет для детектора.

Додатково застосовується темпоральна схема обробки, що виконує повний прохід конвеєра лише на ключових кадрах, тоді як на проміжних результати детекції та маски сегментації переносяться засобами онлайн-трекера. Вибір ключових кадрів здійснюється за

перевищенням порогу відхилення поточного значення $I(c)$ від значення, зафіксованого на попередньому ключовому кадрі: повний прохід призначається саме тоді, коли інформативність модальностей суттєво змінюється, а не за жорстким інтервалом. Це дозволяє автоматично збільшувати частоту ключових кадрів у сценах зі швидкою зміною умов (загорання, задимлення, перехід між приміщеннями) і зменшувати її в стабільних сегментах.

Відповідність факторів деградації, інформативності модальностей та обраного режиму передобробки наведено в Таблиці 1.

Таблиця 1. Правила адаптивного вибору режиму передобробки за метрикою інформативності $I(c)$

Фактор деградації	$I(\text{RGB})$	$I(\text{IЧ})$	Режим передобробки
Нічний час, слабе освітлення	0,18 (дуже низька)	0,78 (висока)	Швидка корекція
Дим, задимлення	0,12 (дуже низька)	0,72 (висока)	Реконструкція
Туман, атмосферне розсіювання	0,22 (дуже низька)	0,58 (середня)	Реконструкція
Гарячий фон	0,52 (середня)	0,21 (дуже низька)	Швидка корекція
Гарячий фон із задимленням	0,16 (дуже низька)	0,23 (дуже низька)	Реконструкція
Часткова оклюзія, малий розмір цілі	0,66 (середня)	0,69 (середня)	Без передобробки

Джерело: розроблено авторами

За нічного освітлення RGB-канал втрачає контраст і енергію градієнта, проте зберігає геометрію сцени, а ІЧ-канал працює в найсприятливіших для себе умовах: температурний контраст між людиною та охолодженим фоном максимальний. Оскільки одна з модальностей має високу інформативність, витратити обчислювальний ресурс на нейромережеве відновлення недоцільно – достатньо швидкої корекції видимого каналу, щоб він залишався корисним джерелом контурних ознак для розділення близько розташованих цілей.

За задимлення ситуація інша: показники серпанку та контрасту в RGB-каналі падають одночасно, і швидка корекція лише підсилює шум, не відновлюючи втраченої інформації, оскільки сигнал спотворено мультиплікативно, а не адитивно. Саме тут виправдана реконструкція RGB-каналу: вона відновлює частину структурної інформації, необхідної для локалізації малорозмірних цілей.

Хоча в RGB-каналі туман і дим спричиняють подібну за механізмом деградацію через розсіювання світла, у тепловому діапазоні їхній вплив різниться: дим є для довгохвильового ІЧ-випромінювання майже прозорим, тоді як щільний туман із крапельною фазою частково його поглинає. Тому за туману інформативність ІЧ-гілки знижується до середнього рівня, а компенсаторна роль реконструкції RGB-каналу зростає.

Ситуації що описані в таблиці терміном «гарячий фон» є показовим винятком: тут деградує саме та модальність, яка зазвичай вважається найнадійнішою. Температура горючих поверхонь перевищує температуру тіла людини, тепловий контраст інвертується. Значна частка пікселів виходить за верхню межу очікуваного діапазону теплової сигнатури людини, компонента $f_6(IЧ)$ падає й потягує за собою інтегральний показник $I(IЧ)$, попри те що локальний контраст і енергія градієнта на гарячих кромках зростають. Хоча й пожежі в більшості випадків супроводжуються сильним задимленням, у випадку коли задимлення буде низьке, або відсутнє, RGB-канал зберігатиме середню інформативність та може навіть отримати перевагу за рахунок додаткового освітлення від полум'я. Реконструкція в цьому випадку не потрібна, оскільки інверсія теплового контрасту не є втратою інформації, яку здатне повернути відновлення: сигнал не пригнічений шумом чи розсіюванням, а змінив знак, натомість потрібне перерозподілення ваг на користь видимого каналу, тому обрано швидку корекцію ІЧ-каналу.

Але пожежа часто створює як «гарячий фон», так і сильне задимлення. Це найскладніший випадок: обидві модальності деградують одночасно, але за незалежними механізмами – дим розсіює світло у видимому діапазоні й обвалює RGB-канал, а інвертований тепловий контраст обвалює ІЧ-канал. Тому вмикається реконструкція RGB-каналу зі швидкою корекцією ІЧ-каналу, режимом кадру стає реконструкція як найважчий із двох, а вагова схема спирається на гілку з дещо вищим показником, доповнюючи її темпоральним контекстом трекінгу.

Останній описаний в таблиці випадок відповідає ситуації, коли обидві модальності інформативні, а джерелом складності є не деградація зображення, а геометрія цілі – малий розмір і часткове перекриття. Передобробка тут не додає інформації, а лише витрачає час, тому вона пропускається; складність компенсується на рівні навчальної вибірки та функції втрат.

Загальне правило, що впливає з таблиці, формулюється у два кроки. Спершу визначається, які канали взагалі потребують обробки: передобробка застосовується лише до тих, чий показник не досягає рівня «середня», тобто $I(c) < 0,45$; якщо цій умові не відповідає жоден канал, передобробка пропускається. Далі для кожного такого каналу режим обирається не за самим значенням $I(c)$, а за тим, яка з компонент виразу (1) спричинила його падіння. Якщо визначальним є зниження оберненої оцінки щільності серпанку за темновим каналом, або відношення сигнал/шум, деградація має розсіювальну, мультиплікативну природу: структурна інформація пригнічена, але не знищена, і частину її можливо відновити – вмикається реконструкція. Якщо падіння показника спричинене компонентами контрасту, ентропії та енергії градієнта, деградація є тональною: сигнал ослаблений, але не спотворений, і достатньо швидкої корекції. Окремо стоїть інверсія теплового контрасту: компоненти контрасту, ентропії та градієнта у цьому разі не знижуються, проте компонента f_0 падає через вихід яскравостей за межі очікуваного діапазону теплової сигнатури людини; відновленню такий сигнал не піддається, тому також обирається швидка корекція. Режимом кадру вважається найважчий із режимів, обраних для окремих каналів.

Часовий бюджет кадру моделюється як сума чотирьох доданків: обчислення метрики (1), передобробки, прямого проходу детектора-сегментатора та модуля крос-модальної уваги. Перший доданок є найдешевшим: усі компоненти показника обчислюються за одним проходом по зменшеній копії кадру й не потребують операцій згортки. Другий доданок є єдиним змінним: він дорівнює нулю в режимі пропуску, залишається малим за швидкої корекції та стає найважчим у режимі реконструкції. Отже, середня вартість кадру визначається не піковою складністю конвеєра, а часткою кадрів послідовності, для яких вмикається реконструкція. Ця частка задається порогами метрики та профілем її компонент, а отже, є керованою: підвищення порогу знижує середнє навантаження ціною повноти виявлення в найважчих кадрах, причому такий компроміс є явним, на відміну від схем із фіксованою передобробкою, де він прихований у конфігурації системи. Додатково вартість кадру зменшується пропорційно до частки ключових кадрів: якщо на послідовність із N кадрів припадає лише $K < N$ ключових, повний прохід виконується в середньому K/N разів на кадр, а на решті витрачається лише вартість трекінгу. Числові значення доданків залежать від платформи та реалізації модулів і належать до предмета експериментальної перевірки.

Наукова новизна дослідження полягає в тому, що, на відміну від існуючих підходів, у яких злиття модальностей виконується зі статичними вагами, а попередня обробка застосовується однаково до всіх кадрів послідовності, удосконалено спосіб детекції та сегментації людей у мультиспектральному відеопотоці, який відрізняється введенням задачно-специфічної кількісної метрики інформативності модальності, адаптивним керуванням глибиною передобробки за профілем її компонент та узгодженою з фізикою деградацій стратегією формування навчальної вибірки. Це дозволяє підвищити очікувану точність виявлення малорозмірних і частково перекритих цілей в умовах низької видимості без втрати швидкодії.

Експериментальну перевірку запропонованого способу здійснено на відкритих мультиспектральних наборах даних KAIST Multispectral Pedestrian, LLVIP та FMB, об'єднаних у відтворюваний експериментальний протокол MS-ER (Multispectral person detection & segmentation in Emergency-like conditions). Матеріали для відтворення експериментальної вибірки, уніфіковані анотації, фіксовані списки навчальних і тестових вибірок, контрольні суми та скрипти підготовки даних розміщено у відкритому репозиторії GitHub broleh/ms-er-dataset, доступному за посиланням: <https://github.com/broleh/ms-er-dataset> [18]. Вихідні RGB- та ІЧ-зображення завантажуються з офіційних джерел відповідних наборів даних згідно з умовами їх ліцензування.

Для основної перевірки детекції використано підмножину KAIST із 9853 пар попарно узгоджених RGB- та тепловізійних кадрів, з яких 7601 пару віднесено до навчальної вибірки, а 2252 – до тестової. KAIST містить синхронізовані видимі та теплові зображення, отримані за денних і нічних умов, та анотації пішоходів; у MS-ER для нього передбачено аналіз за масштабом об'єкта, ступенем перекриття та часом доби. В анотаціях додатково зберігаються ознаки *occlusion*, *scale*, *reasonable* та належність кадру до відеопослідовності, що дає змогу окремо аналізувати ближні, середні й віддалені об'єкти та частково перекритих людей.

Другим незалежним набором для перевірки детекції обрано LLVIP, що містить 15 488 просторово й часово узгоджених пар RGB–ІЧ зображень. У використаному стандартному розподілі 12 025 пар належать до навчальної вибірки, а 3463 – до тестової. Особливістю LLVIP є значна частка сцен із низькою та дуже низькою освітленістю, тому його використано для оцінювання здатності способу компенсувати зниження інформативності RGB-каналу за рахунок теплової модальності. Оригінальний набір містить 30 976 зображень, тобто 15 488 пар, які строго вирівняні у просторі та часі, а пішоходи мають відповідну розмітку.

Для експериментальної перевірки сегментаційної складової використано FMB [19]. У сформованій вибірці враховано 1500 пар RGB–ІЧ кадрів, з яких 1220 використано для навчання та 280 – для тестування. На відміну від KAIST і LLVIP, ця частина протоколу застосовувалася передусім для оцінювання піксельної сегментації людей за масками. У структурі MS-ER для FMB зберігаються шляхи до видимого, інфрачервоного зображення та карти класів, причому маска людини формується за відповідною міткою класу. Наявність сцен із туманом, дощем і низькою освітленістю дозволяє використовувати цю частину вибірки для перевірки роботи сегментаційної голови за погіршених умов спостереження.

В результаті сформована експериментальна база охоплювала 26 841 пару RGB–ІЧ кадрів, або 53 682 окремих зображення. Для навчання сегментаційної голови додатково було сформовано 55 427 масок людей. Для дослідження стійкості передбачено також окремий етап синтетичної деградації кадрів, що включає дим, туман, шум, розмиття, недостатнє освітлення, гарячий фон і пожежоподібні візуальні умови. Ця частина не замінювала тестування на реальних відкритих наборах, а використовувалася як додатковий *robustness*-тест після базової перевірки на KAIST, LLVIP та FMB. Саме таке розмежування важливе, оскільки дозволяє окремо показати результати на реальних відкритих даних і результати штучного стрес-тестування.

Відтворюваність експерименту забезпечувалася фіксацією розподілу даних у каталозі *splits*, уніфікованими анотаціями, SHA-256 контрольними сумами використаних зображень і сценаріями автоматизованої побудови вибірки. Перед проведенням експериментів процедура верифікації повторно обчислювала контрольні суми зображень і перевіряла відповідність кількості файлів еталонним значенням. Таким чином, використані тестові підмножини можуть бути однозначно відтворені незалежним дослідником із вихідних відкритих наборів даних.

В Таблиці 2 наведено показники точності детекції та сегментації людей для досліджуваних конфігурацій.

З Таблиці 2 випливає, що поєднання RGB та ІЧ-даних з адаптивним злиттям ознак і попередньою обробкою забезпечує стабільніше виявлення та точніше виділення контурів

людей порівняно з одномодалними й статичними конфігураціями. Це підтверджує доцільність запропонованого підходу для мультиспектрального відеопотоку в складних умовах надзвичайних ситуацій.

Таблиця 2. Показники точності детекції та сегментації людей для досліджуваних конфігурацій

Показник	M ₁ RGB-only	M ₂ Thermal-only	M ₃ Static RGB+ІЧ	M ₄ ICAFusion [8]	M ₅ Без адаптивної передобробки	M ₆ запропонована модель
mAP@0.5, %	90,8 ± 0,7	94,5 ± 0,5	96,7 ± 0,4	98,4 ± 0,3	98,1 ± 0,3	98,7 ± 0,2
mAP@0.5:0.95, %	50,2 ± 0,8	61,6 ± 0,7	63,4 ± 0,6	64,5 ± 0,5	68,2 ± 0,5	70,6 ± 0,5
Precision, %	90,4 ± 0,6	92,1 ± 0,5	93,8 ± 0,5	95,1 ± 0,4	95,8 ± 0,4	96,4 ± 0,3
Recall, %	85,8 ± 0,9	89,7 ± 0,7	91,5 ± 0,6	93,0 ± 0,5	94,1 ± 0,5	95,3 ± 0,4
F1, %	88,0 ± 0,7	90,9 ± 0,6	92,6 ± 0,5	94,0 ± 0,4	94,9 ± 0,4	95,8 ± 0,3
mIoU, %	66,3 ± 0,8	70,8 ± 0,7	73,5 ± 0,7	75,1 ± 0,6	77,6 ± 0,6	80,4 ± 0,6

Джерело: розроблено авторами

Примітка. Значення наведено як середнє ± стандартне відхилення за п'ятьма незалежними запусками (n=5)

В Таблиці 3 наведені результати дослідження сегментації та детекції людей у мультиспектральному відеопотоці в умовах надзвичайних ситуацій.

Таблиця 3. Результати дослідження сегментації та детекції людей у мультиспектральному відеопотоці в умовах надзвичайних ситуацій

Показник	M ₁ RGB-only	M ₂ Thermal-only	M ₃ Static RGB+ІЧ	M ₄ ICAFusion [8]	M ₅ Без адаптивної передобробки	M ₆ Запропонована модель
AP _S , %	43,8 ± 1,1	52,7 ± 0,9	57,8 ± 0,8	61,9 ± 0,7	64,8 ± 0,7	68,5 ± 0,7
Recall _S , %	61,7 ± 1,2	70,8 ± 1,0	76,4 ± 0,9	80,2 ± 0,8	83,5 ± 0,7	87,1 ± 0,6
F1 _S , %	66,2 ± 1,0	73,5 ± 0,9	78,9 ± 0,8	82,1 ± 0,7	85,1 ± 0,6	88,4 ± 0,6
AP _{occ} , %	46,1 ± 1,0	54,9 ± 0,9	60,2 ± 0,9	63,5 ± 0,8	67,1 ± 0,8	71,6 ± 0,8
Recall _{occ} , %	64,3 ± 1,1	72,4 ± 1,0	78,1 ± 0,8	81,4 ± 0,8	84,2 ± 0,7	88,3 ± 0,6
F1 _{occ} , %	68,5 ± 1,0	75,1 ± 0,9	80,3 ± 0,8	83,2 ± 0,7	85,9 ± 0,6	89,2 ± 0,5
mIoU _{occ} , %	55,8 ± 1,1	61,7 ± 0,9	66,4 ± 0,8	69,8 ± 0,7	73,2 ± 0,7	77,5 ± 0,6
95 % CI mAP@0.5:0.95, %	49,4–51,0	60,9–62,3	62,7–64,1	63,9–65,1	67,6–68,8	70,0–71,2

Джерело: розроблено авторами

Дані з Таблиці 3 ілюструють, що запропонована модель M₆ забезпечує найкращі результати саме для складних умов надзвичайних ситуацій. Для малорозмірних людей досягнуто AP_S = 68,5 %, Recall_S = 87,1 % та F1_S = 88,4 %, а для частково перекритих людей – AP_{occ} = 71,6 %, Recall_{occ} = 88,3 %, F1_{occ} = 89,2 % і mIoU_{occ} = 77,5 %. Порівняно з ICAFusion [8] запропонована модель демонструє вищу точність виявлення та сегментації малих і частково перекритих об'єктів, що підтверджує ефективність адаптивної обробки RGB та ІЧ-даних у складних сценах.

За результатами експериментальної перевірки системи детекції та сегментації людей у мультиспектральному RGB–ІЧ відеопотоці встановлено, що запропонована система забезпечує одночасну детекцію та сегментацію людей за RGB- та ІЧ-даними й адаптивно застосовує різні режими передобробки залежно від складності кадру. Під час обробки 5000 кадрів середня продуктивність становила 38,0 FPS, максимальна – 49,8 FPS, а пікове використання GPU-пам'яті – 6,4 ГБ. Без додаткової передобробки опрацьовано 46,8 % кадрів, швидку корекцію застосовано для 31,6 %, а глибоку реконструкцію – для 21,6 %, що підтверджує можливість адаптивної обробки мультиспектрального відеопотоку в режимі, близькому до реального часу.

В Таблиці 4 наведено результати моделювання швидкодії запропонованого способу за різних режимів передобробки (Апаратна платформа: NVIDIA GeForce RTX 3090 24 ГБ, Intel Core i9-12900K, 32 ГБ RAM; вхідні RGB/ІЧ-кадри – 640 × 640 px; batch size – 1).

Таблиця 4. Результати моделювання швидкодії запропонованого способу за різних режимів передобробки

Режим передобробки	Частка кадрів, %	Повний прохід, мс/кадр	Трекінг, мс/кадр	Реконструкція / корекція, мс/кадр	Середня частота, кадр/с	Гранична частота, кадр/с	GPU-пам'ять, ГБ
Без передобробки	46,8	21,7	2,4	0,0	46,1	49,8	5,7
Швидка корекція	31,6	25,9	2,5	3,8	38,6	42,1	5,9
Поглиблена реконструкція	21,6	38,4	2,6	15,9	26,0	29,4	6,4
Загальний результат	100,0	26,3	2,5	4,6	38,0	49,8	6,4*

Джерело: розроблено авторами

* Для загального результату вказано пікове споживання GPU-пам'яті

Із Таблиці 4 випливає, що моделювання швидкодії показало залежність обчислювальних витрат від активованого режиму передобробки. Для 46,8 % кадрів додаткова передобробка не застосовувалася, при цьому повний час проходу становив 21,7 мс, а середня продуктивність – 46,1 кадр/с. Швидка корекція активувалася для 31,6 % кадрів і збільшувала час обробки до 25,9 мс, що відповідало 38,6 кадр/с. Найбільші витрати отримано для поглибленої реконструкції, яка застосовувалася до 21,6 % кадрів: повний час проходу зростав до 38,4 мс, а продуктивність знижувалася до 26,0 кадр/с.

У середньому для всього відеопотоку час повної обробки становив 26,3 мс на кадр при продуктивності 38,0 кадр/с. Час трекінгу залишався відносно стабільним і знаходився в межах 2,4–2,6 мс, тоді як найбільший додатковий внесок у затримку формувала реконструкція – до 15,9 мс. Споживання GPU-пам'яті змінювалося від 5,7 до 6,4 ГБ залежно від режиму. Таким чином, обчислювально найбільш витратний режим застосовувався лише приблизно до п'ятої частини кадрів, що в модельному бенчмарку дозволило зберегти середню продуктивність на рівні близько 38 кадр/с.

Висновки. У результаті виконаного дослідження одержано такі результати:

1. Було розроблено та експериментально перевірено спосіб сегментації й детекції людей у мультиспектральному RGB–ІЧ відеопотоці, орієнтований на роботу за змінної видимості та характерних для надзвичайних ситуацій деградацій. Запропонований підхід забезпечує динамічне використання інформації двох спектральних каналів відповідно до їх поточного стану замість фіксованого внеску модальностей.

2. Експериментальна перевірка на відкритих мультиспектральних наборах KAIST, LLVIP та FMB показала перевагу повної конфігурації M_6 над одномодальними моделями, статичним RGB–ІЧ-злиттям та ICAFusion за показниками детекції й сегментації. Особливо помітний ефект отримано для складних цілей: для малорозмірних людей досягнуто $AP_S = 68,5\%$, $Recall_S = 87,1\%$ і $F1_S = 88,4\%$, а для частково перекритих — $AP_{occ} = 71,6\%$, $Recall_{occ} = 88,3\%$, $F1_{occ} = 89,2\%$ та $mIoU_{occ} = 77,5\%$. Це підтверджує доцільність локального перерозподілу внеску RGB- та ІЧ-ознак і використання сегментаційної гілки за дефіциту візуальних ознак.

3. Установлено, що внесок спектральних каналів суттєво змінюється залежно від характеру сцени. За слабого освітлення та задимлення перевагу має ІЧ-канал із показниками інформативності 0,78 та 0,72 відповідно, тоді як за гарячого фону інформативність RGB-каналу становить 0,52 проти 0,21 для ІЧ. За одночасного гарячого фону та задимлення показники обох каналів знижуються до 0,16 і 0,23, що обґрунтовує перехід до більш глибокої передобробки. Таким чином, кількісне оцінювання інформативності дає змогу узгоджувати

спосіб обробки кадру з фактичним типом деградації, а не застосовувати однакову процедуру до всього відеопотоку.

4. Підтверджено також обчислювальну доцільність адаптивної схеми. Для 46,8 % кадрів додаткова передобробка не була потрібна, для 31,6 % застосовувалася швидка корекція і лише для 21,6 % — поглиблена реконструкція. На NVIDIA GeForce RTX 3090 за batch size = 1 середній час повної обробки становив 26,3 мс на кадр, що відповідало 38,0 кадр/с, а пікове використання GPU-пам'яті становило 6,4 ГБ. Отже, обчислювально найдорожча процедура застосовується вибірково, що дозволяє зберігати продуктивність, достатню для обробки відеопотоку в реальному часі.

5. Отримані результати підтверджують досягнення поставленої мети: поєднання оцінювання інформативності модальностей, адаптивної передобробки, крос-модального злиття та спеціалізованої роботи з малорозмірними й частково перекритими цілями підвищує стійкість детекції та сегментації за складних умов без втрати необхідної швидкодії. Практичне значення результатів полягає у можливості використання запропонованого підходу в мультиспектральних системах відеоспостереження та пошуково-рятувальних комплексах, де умови спостереження можуть істотно змінюватися безпосередньо в межах однієї відеопослідовності.

Перспективою подальших досліджень є перевірка перенесуваності встановлених порогів інформативності на інші сенсорні комплекси, розширення експериментів на реальні відеопослідовності пожеж і пошуково-рятувальних операцій та дослідження можливості зменшення обчислювальних витрат на вбудованих апаратних платформах.

Конфлікт інтересів: Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті

Фінансування: Дослідження проводилося без фінансової підтримки

Доступність даних: Матеріали для відтворення експериментальної вибірки (уніфіковані анотації, фіксовані розбиття на навчальну та тестову вибірки, контрольні суми SHA-256 і скрипти підготовки даних) розміщено у відкритому репозиторії <https://github.com/broleh/ms-er-dataset> [18]. Вихідні RGB- та ІЧ-зображення наборів KAIST Multispectral Pedestrian, LLVIP та FMB доступні з офіційних джерел відповідних наборів даних згідно з умовами їх ліцензування. Числові результати експериментів наведено в основному тексті рукопису

Використання засобів штучного інтелекту (ШІ): Під час підготовки рукопису використовувався інструмент Claude (Anthropic, модель Claude Opus 5) штучного інтелекту (ШІ) для граматичної та стилістичної перевірки тексту. Усі фрагменти, опрацьовані за його допомогою, були перевірені й відредаговані автором. ШІ не використовувався для генерації тексту, формул і таблиць, формування наукових положень, отримання результатів або формулювання висновків. Наукові положення, результати та висновки є власним інтелектуальним внеском авторів

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Li, S., Tang, H. “Multimodal alignment and fusion: A survey”. *International Journal of Computer Vision*. 2026; 134 (3): 103. DOI: <https://doi.org/10.1007/s11263-025-02667-1>.
2. Akkaya, I. B., Altinel, F. & Halici, U. “Self-training guided adversarial domain adaptation for thermal imagery”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021. p. 4322–4331. DOI: <https://doi.org/10.1109/CVPRW53098.2021.00488>.
3. Devaguptapu, C., Akolekar, N., Sharma, M. M. & Balasubramanian, V. N. “Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019. p. 1029–1038. DOI: <https://doi.org/10.1109/CVPRW.2019.00135>.
4. Hwang, S., Park, J., Kim, N., Choi, Y. & Kweon, I. S. “Multispectral pedestrian detection: Benchmark dataset and baseline”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015. p. 1037–1045. DOI: <https://doi.org/10.1109/CVPR.2015.7298706>.
5. Kieu, M., Bagdanov, A. D., Bertini, M. & Del Bimbo, A. “Task-conditioned domain adaptation for pedestrian detection in thermal imagery”. In *Proceedings of the European*

Conference on Computer Vision (ECCV). 2020. p. 546–562. DOI: https://doi.org/10.1007/978-3-030-58542-6_33.

6. Jia, X., Zhu, C., Li, M., Tang, W. & Zhou, W. “LLVIP: A visible-infrared paired dataset for low-light vision”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2021. p. 3496–3504. DOI: <https://doi.org/10.1109/ICCVW54120.2021.00389>.

7. Ghari, B., Tourani, A., Shahbahrami, A. & Gaydadjiev, G. “Pedestrian detection in low-light conditions: A comprehensive survey”. *Image and Vision Computing*. 2024; 148: 105106. DOI: <https://doi.org/10.1016/j.imavis.2024.105106>.

8. Shen, J., Chen, Y., Liu, Y., Zuo, X., Fan, H. & Yang, W. “ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection”. *Pattern Recognition*. 2024; 145: 109913, <https://scopus.com/pages/publications/85170537683>.

DOI: <https://doi.org/10.1016/j.patcog.2023.109913>.

9. Li, C., Song, D., Tong, R. & Tang, M. “Illumination-aware Faster R-CNN for robust multispectral pedestrian detection”. *Pattern Recognition*. 2019; 85: 161–171, <https://scopus.com/pages/publications/85051983446>.

DOI: <https://doi.org/10.1016/j.patcog.2018.08.005>.

10. Guan, D., Cao, Y., Yang, J., Cao, Y. & Yang, M. Y. “Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection”. *Information Fusion*. 2019; 50: 148–157. DOI: <https://doi.org/10.1016/j.inffus.2018.11.017>.

11. Zhou, K., Chen, L. & Cao, X. “Improving multispectral pedestrian detection by addressing modality imbalance problems”. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020. p. 787–803. DOI: https://doi.org/10.1007/978-3-030-58523-5_46.

12. Tang, H., Zhu, H., Fei, L., Wang, T., Cao, Y. & Xie, C. “Low-illumination image enhancement based on deep learning techniques: A brief review”. *Photonics*. 2023; 10 (2): 198. DOI: <https://doi.org/10.3390/photonics10020198>.

13. Gui, J., Cong, X., Cao, Y., Ren, W., Zhang, J., Zhang, J., Cao, J. & Tao, D. “A comprehensive survey and taxonomy on single image dehazing based on deep learning”. *ACM Computing Surveys*. 2023; 55 (13s): 279. DOI: <https://doi.org/10.1145/3576918>.

14. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. “Microsoft COCO: Common objects in context”. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 2014. p. 740–755. DOI: https://doi.org/10.1007/978-3-319-10602-1_48.

15. He, K., Sun, J. & Tang, X. “Single image haze removal using dark channel prior”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2011; 33 (12): 2341–2353. DOI: <https://doi.org/10.1109/TPAMI.2010.168>.

16. Jocher, G., Chaurasia, A. & Qiu, J. “Ultralytics YOLOv8”. *GitHub*. 2023. – Available from: <https://github.com/ultralytics/ultralytics>.

17. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. “Focal loss for dense object detection”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020; 42 (2): 318–327. DOI: <https://doi.org/10.1109/TPAMI.2018.2858826>.

18. Uhryn, D., Breslavskiy, O. “MS-ER: Multispectral person detection & segmentation in emergency-like conditions”. *GitHub*. 2026. – Available from: <https://github.com/broleh/ms-er-dataset>.

19. Liu, J., Liu, Z., Wu, G., Ma, L., Liu, R., Zhong, W., Luo, Z. & Fan, X. “Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation”. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023. p. 8115–8124. DOI: <https://doi.org/10.1109/ICCV51070.2023.00745>.

Отримано 27.08.2026

Отримано після перегляду 25.09.2026

Прийнято 05.10.2026

DOI: <https://doi.org/10.15276/ict.03.2026.55>

UDC 004.932:004.85

Segmentation and detection of humans in a multispectral video stream under emergency conditions

Dmytro I. Uhryn¹⁾

ORCID: <https://orcid.org/0000-0003-4858-4511>; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300

Oleh I. Breslavskyi¹⁾

ORCID: <https://orcid.org/0009-0005-5164-2913>; breslavskyi.oleh@chnu.edu.ua

¹⁾Yuriy Fedkovych Chernivtsi National University, 2 Kotsiubynskyi Str. Chernivtsi, 58012, Ukraine

ABSTRACT

This work focuses on improving the accuracy and robustness of human detection and segmentation in multispectral video streams during emergency situations, while maintaining performance levels sufficient for real-time operation. It is demonstrated that directly transferring detectors trained on visible-spectrum images to the infrared spectrum is ineffective due to cross-modal domain shift; the modalities differ in the physical nature of the signal, noise levels, contrast characteristics, and the set of available textural features. The study analyzes key approaches to human detection and segmentation in the infrared range, including thermal detectors and specialized datasets, domain adaptation and feature transfer from the visible spectrum, multispectral detectors with varying fusion levels, and methods for compensating for modal imbalance. A common limitation of these approaches is identified: fixed reliance on specific modalities and a uniform level of preprocessing depth across all frames. An improved method for human detection and segmentation is proposed, based on a dual-branch YOLOv8m-seg architecture that integrates a quantitative metric of modality informativeness, adaptive frame preprocessing, and cross-modal feature fusion with dynamic weighting. The informativeness of RGB and IR channels is determined based on normalized characteristics—specifically local root-mean-square contrast, luminance distribution entropy, Tenengrad gradient energy, an inverse haze density estimate, and the signal-to-noise ratio—with the thermal channel additionally accounting for target thermal range consistency. Metric values fall within the [0, 1] range and are categorized into four levels: very low ($I < 0.25$), low ($0.25 \leq I < 0.45$), medium ($0.45 \leq I < 0.70$), and high ($I \geq 0.70$). When $I < 0.45$, the intervention mode is determined by the profile of the metric components: rapid correction or neural network-based restoration is applied depending on the nature of the degradation, whereas additional preprocessing is bypassed if informativeness is sufficient. Cross-modal fusion is implemented using variable spatial weights for RGB and IR features, determined according to the local informativeness of each modality. Additionally, a training dataset generation strategy aligned with the physics of degradation is employed, alongside techniques for handling small-scale and partially occluded targets. Experimental validation was conducted using the KAIST Multispectral Pedestrian, LLVIP, and FMB open multispectral datasets, consolidated into the reproducible MS-ER protocol; materials for experiment replication are available in a public GitHub repository. The experimental dataset comprised 26,841 pairs of RGB-IR frames (totaling 53,682 images), with an additional 55,427 human masks generated for training the segmentation component. Six configurations (M1–M6) were used for a controlled comparison: RGB-only, thermal-only, static RGB-IR fusion, ICAFusion, the proposed model without adaptive preprocessing, and the full M6 model featuring an informativeness metric, adaptive preprocessing, and local modality weight control. An experimental run of 5,000 frames revealed that 46.8% of frames were processed without additional preprocessing, 31.6% underwent rapid correction, and 21.6% required deep reconstruction; the average processing time was 26.3 ms/frame, with a mean throughput of 38.0 FPS (peaking at 49.8 FPS) and peak GPU memory usage of 6.4 GB. The results confirm the viability of the proposed method and the feasibility of adaptively redistributing computational resources based on the real-time informativeness of the RGB and IR channels. Applying the most resource-intensive mode only to degraded frames maintains sufficient performance for real-time multispectral video processing while leveraging the complementary information from RGB and IR modalities for human detection and segmentation under challenging surveillance conditions.

Keywords: multispectral video stream; human detection; semantic segmentation; infrared range; cross-modal domain shift; cross-modal feature fusion; modality-informativeness metric; adaptive preprocessing; small-scale targets; real-time processing

ІНФОРМАЦІЯ ПРО АВТОРІВ



Угрин Дмитро Ілліч - доктор технічних наук, професор, професор кафедри Комп'ютерних наук. Чернівецький національний університет імені Юрія Федьковича, вул. Коцюбинського, 2. Чернівці, 58012, Україна
Галузь досліджень: інтелектуальний аналіз даних, інформаційні технології підтримки прийняття рішень, системи ройового інтелекту, галузеві геоінформаційні системи

Dmytro I. Uhryn - Doctor of Engineering Sciences, Professor, Department of Computer Science. Yuriy Fedkovych Chernivtsi National University, 2, Kotsiubynskyi Str. Chernivtsi, 58012, Ukraine
ORCID: <https://orcid.org/0000-0003-4858-4511>; d.ugryn@chnu.edu.ua. Scopus Author ID: 57163746300

Research field: data mining, decision support information technologies, swarm intelligence systems, industry-specific geographic information systems



Бреславський Олег Ігорович - аспірант кафедри Програмного забезпечення комп'ютерних систем. Чернівецький національний університет імені Юрія Федьковича, вул. Коцюбинського, 2. Чернівці, 58012, Україна
Галузь досліджень: комп'ютерний зір, глибинне навчання, вбудовані системи, відеоаналітика

Oleh I. Breslavskyi - PhD student, Department of Computer Systems Software. Yuriy Fedkovych Chernivtsi National University, 2, Kotsiubynskyi Str. Chernivtsi, 58012, Ukraine
ORCID: <https://orcid.org/0009-0005-5164-2913>; breslavskyi.oleh@chnu.edu.ua

Research field: computer vision, deep learning, embedded systems, video analytics