



СЦЕНОЗАЛЕЖНЕ АНСАМБЛЕВЕ ВИЯВЛЕННЯ ОБ'ЄКТІВ НА АЕРОЗНІМКАХ БЕЗПІЛОТНИКІВ: МЕЖІ КОРИСНОСТІ МАРШРУТИЗАЦІЇ КАДРІВ ЗА СЦЕНАМИ

Досліджено сценозалежне ансамблеве виявлення об'єктів на аерознімках безпілотних літальних апаратів (БПЛА), у якому класифікатор сцени (день/ніч/дощ) спрямовує кадр до спеціалізованого пер-сценового ансамблю детекторів зі злиттям передбачень методом зваженого злиття рамок WBF (англ. *Weighted Boxes Fusion*). Проведено ізолювальний аналіз, який відокремлює внесок сценозалежної маршрутизації даних від внеску самого ансамблювання детекторів; нової архітектури детектора не запропоновано. Виявлено та виправлено систематичні дефекти протоколу оцінювання детекторів – використання порогу впевненості розгортання системи замість оцінювального, невідповідність розмітки 10 і 8 класів та обрізання постоброблення за часом, – які спотворюють оцінки окремих моделей на десятки відсотків і роблять порівняння архітектур детекторів ненадійним. Встановлено, що за коректним протоколом ансамбль детекторів WBF трьох детекторів стабільно перевершує найкращу одиночну модель у кожній сцені (близько $+0,05 \text{ mAP}@50$) ціною приблизно триразового зниження швидкості інференсу, тобто йдеться про компроміс Парето. З'ясовано, що сценозалежна маршрутизація кадру не перевершує єдиний сценонезалежний ансамбль детекторів тих самих моделей (різниця $-0,0006 \text{ mAP}@50$), і цей результат відтворено на гетерогенному пулі з п'яти поколінь YOLO. Показано, що надлишковість передбачень пулу є властивістю навчального розподілу даних, а не архітектури детектора, у межах перевіреного простору публічних даних, доступних станом на липень 2026 року. Встановлено, що на другому наборі даних UAVDT (англ. *Unmanned Aerial Vehicle Benchmark: Detection and Tracking*) з офіційними мітками умов (день/ніч/туман) маршрутизація кадру вперше дає статистично значущий вииграш за фіксованого обчислювального бюджету ансамблю ($+0,014 \text{ mAP}@50$), проте злиття всього пулу моделей без відбору знову поглинає цю перевагу. Сформульовано і підтверджено на двох наборах даних уточнену умову корисності маршрутизації даних: сценоспецифічна перевага має бути невідновлюваною злиттям повного пулу. Показано, що практичною нішею маршрутизації даних є жорсткий обчислювальний бюджет за доменного зсуву з екзогенним джерелом знання про сцену, оскільки класифікатор за виглядом кадру розпізнає критичні несприятливі умови (туман) тільки у 32 % випадків.

Ключові слова: зважене злиття рамок; глибоке навчання моделей; згорткові нейронні мережі; еталонний набір даних VisDrone; відкритий набір даних UAVDT; доменний зсув.

Вступ / Introduction

Виявлення об'єктів на аерознімках безпілотних літальних апаратів (БПЛА) – складне завдання комп'ютерного зору з прикладною цінністю для моніторингу, безпеки та аналізу трафіку. Еталонний набір даних VisDrone-DET (англ. *Vision Meets Drones*) [22] унаочнює три головні труднощі: дрібність об'єктів (близько 54 % екземплярів менші за 32×32 пікселі), високу щільність об'єктів (сотні на кадр) і варіативність умов знімання (день, ніч, дощ, туман). Остання означає, що одна універсальна модель працює нерівномірно: розрив між сценами для найкращої одиночної моделі сягає $0,544 \text{ mAP}@50$ удень проти $0,334$ уночі, тобто падіння якості виявлення об'єктів приблизно на третину. Це мотивує

сценозалежний підхід до побудови детекційної системи.

Центральна гіпотеза сценозалежного підходу полягає в тому, що маршрутизація кадрів за сценою до спеціалізованого пер-сценового ансамблю детекторів зі спеціально підібраними вагами злиття перевершує і найкращу одиночну модель, і єдиний глобальний (сценонезалежний) ансамбль детекторів, а архітектурна різноманітність детекторів посилює вииграш. Ця гіпотеза виглядає правдоподібною: різні сцени висувають різні вимоги до детектора, тому природно очікувати, що спеціалізовані конфігурації ансамблю матимуть перевагу над універсальною. Утім, у наявній літературі бракує чесного ізолювального аналізу сценозалежних детекційних систем, який відокремлював би внесок ансамблювання

© Copyright 2026 by the author(s)

Руда Христина Степанівна, д-р філософії, доцент, кафедра захисту інформації. Email: Khrystyna.S.Ruda@lpnu.ua;

<https://orcid.org/0000-0001-8644-411X>

Кромкач Владислав Олександрович, асистент, кафедра захисту інформації. Email: vlad.kromkach@gmail.com;

<https://orcid.org/0009-0001-5608-5715>

Цитування за ДСТУ: Руда Х. С., Кромкач В. О. Сценозалежне ансамблеве виявлення об'єктів на аерознімках безпілотників: межі корисності маршрутизації кадрів за сценами. Науковий вісник НЛТУ України. 2026, т. 36, № 4. С. 166–177.

Citation APA: Ruda, Kh. S., & Kromkach, V. O. (2026). Scene-dependent ensemble object detection on uav aerial imagery: the limits of usefulness of scene-based frame routing. *Scientific Bulletin of UNFU*, 36(4), 166–177. <https://doi.org/10.36930/40360419>

детекторів від внеску маршрутизації даних разом із кількісним поясненням механізму їхньої взаємодії через кореляцію передбачень моделей.

Потрібно одразу окреслити жанр цієї наукової роботи. Це ізолювальне дослідження, в якому не пропонуємо нової архітектури детектора чи нового методу злиття; хочемо експериментально відокремити внесок сценозалежної маршрутизації даних від внеску самого ансамблювання детекторів, який у літературі зазвичай демонструється тільки інтегрально. Відповідно, внесок нашої роботи методологічний: коректний протокол оцінювання детекторів, ізолювальні порівняння архітектур зі статистичними гарантіями та сформульована й перевірена на двох наборах даних умова, за якої маршрутизація кадру корисна.

Окремою методологічною проблемою є коректність протоколу оцінювання детекторів. Як показано далі, поширені дефекти – обчислення mAP за порогом впевненості розгортання системи замість оцінювального, невідповідність розмітки класів між донавченими вагами та еталонною анотацією, обрізання постоброблення за часом – здатні спотворити оцінку окремих моделей на десятки відсотків, тож будь-які висновки про переваги архітектур детекторів чи ансамблів без виправлення протоколу є ненадійними.

Об'єкт дослідження – процес автоматизованого виявлення об'єктів на аерознімках БпЛА за змінних умов знімання.

Предмет дослідження – методи ансамблювання детекторів об'єктів та сценозалежної маршрутизації кадрів до спеціалізованих конфігурацій злиття, що дасть змогу кількісно розмежувати внески цих двох механізмів у підсумкову якість виявлення об'єктів.

Мета роботи – розробити методику сценозалежного ансамблевого виявлення об'єктів на аерознімках безпілотників, яка відокремлювала б внесок сценозалежної маршрутизації кадрів від внеску ансамблювання детекторів, що дасть змогу сформулювати та експериментально підтвердити уточнену умову корисності маршрутизації даних, показати зумовленість надлишковості передбачень пулу навчальним розподілом даних, а також кількісно оцінити систематичні дефекти протоколу оцінювання детекторів на щільних аерознімках.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- виявити та виправити систематичні дефекти протоколу оцінювання детекторів на щільних аерознімках, що дасть змогу зробити порівняння моделей та ансамблів детекторів коректним і відтворюваним;
- кількісно оцінити виграш ансамблювання детекторів методом зваженого злиття рамок відносно одиночних моделей у кожній сцені, що забезпечить кількісну основу для вибору між одиночною моделлю та ансамблем детекторів за наявного обчислювального бюджету;
- ізолювати внесок сценозалежної маршрутизації даних від внеску ансамблювання детекторів на гомогенному та гетерогенному пулах моделей, що сприятиме коректному встановленню джерела виграша виявлення об'єктів у сценозалежній системі;
- пояснити отримані ефекти через кореляцію передбачень моделей пулу, що уможливить механістичне пояснення отриманих результатів і прогнозування корисності маршрутизації даних для нових пулів моделей;
- перевірити відтворюваність висновків на другому наборі даних з офіційними мітками умов та оцінити вартість заміни ідеального знання сцени реальним класифікатором,

що дасть можливість окреслити практичну нішу сценозалежної маршрутизації кадрів та вимоги до джерела знання про сцену.

Аналіз останніх досліджень та публікацій. У роботі [14] запропоновано одностадійну парадигму виявлення об'єктів детектором YOLO, яка виконує локалізацію та класифікацію за один прохід мережі й забезпечує високу швидкість оброблення. У наступних поколіннях цю лінію детекторів розвинено зі збереженням загальної парадигми щільних передбачень: YOLOv8 [9] працює в режимі реального часу з надзвичайно високою точністю та швидкістю; YOLOv9 [19] – з програмованою градієнтною інформацією; YOLOv10 [18] – з навчанням без немаксимального пригнічення; Ultralytics YOLO11 [8] – досягає вищої точності розпізнавання (mAP), використовуючи на 22 % менше параметрів, ніж попередня версія YOLOv8, оптимізована для роботи як у хмарі, так і на периферійних (edge) пристроях.

У дослідженні [4] запропоновано трансформерну постановку детекції DETR, яка формулює задачу як передбачення множини та усуває немаксимальне пригнічення і якорі завдяки двочастковому зіставленню; водночас фіксована кількість запитів (близько 100) структурно обмежує цю лінію на щільних сценах, характерних для аерознімків.

У роботі [21] трансформерну парадигму адаптовано до реального часу (RT-DETR) і показано конкурентність із детекторами YOLO за швидкістю та точністю, а у дослідженні [13] у трансформерний детектор повернено немаксимальне пригнічення, що поліпшує збіжність і якість.

У роботі [20] представлено спеціалізовану для аерознімків адаптацію UAV-DETR і повідомлено, що на наборі даних VisDrone базовий RT-DETR досягає близько 0,47 mAP@50, а запропонована авторами адаптація становить 0,49-0,51. Ці значення отримані за 10-класовою розміткою та іншими роздільностями, тож прямо не зіставні з наведеними далі, а публічних донавчених ваг для набору даних VisDrone автори не оприлюднили – саме доступність ваг, а не потенціал архітектури детектора, обмежує наведений далі гейт крос-парадигмальних кандидатів.

У дослідженні [2] запропоновано метод Soft-NMS, який замість відкидання конкурентних рамок плавно знижує їхню впевненість, що зменшує втрати на перекритих об'єктах порівняно із класичним немаксимальним пригніченням.

У роботі [17] розроблено метод зваженого злиття рамок WBF (англ. *Weighted Boxes Fusion*), який не відкидає, а усереднює координати збіжних рамок з вагами за впевненістю, завдяки чому локалізація в ансамблях зазвичай поліпшується; ключовою передумовою ефективності ансамблю автори називають декореляцію помилок його членів – якщо моделі помиляються однаково, злиттю немає чого виправляти.

У праці [7] введено ідею сумішей локальних експертів зі спрямуванням входу до спеціалізованого експерта, а у дослідженні [16] її масштабовано до розріджених архітектур детекторів з умовними обчисленнями (Mixture-of-Experts), де для кожного входу активується тільки частина мережі. Сценозалежна маршрутизація детекторів є перенесенням цієї ідеї на рівень цілих моделей, проте з принциповою відмінністю: у МоЕ вентильна функція навчається спільно з експертами під остаточну

функцію втрат, тоді як тут вентиль (класифікатор сцени) фіксований і зовнішній щодо детекторів, а "експерти" – незалежно донавчені моделі; результати цієї роботи стосуються саме другої постановки завдання і не переносяться на першу.

Проведений аналіз останніх досліджень та публікацій показує, що ансамблювання детекторів та умовні обчислення розвиваються як окремі напрями, а роботи зі сценозалежної детекції зазвичай демонструють інтегральний вигравш системи, не відокремлюючи внесок маршрутизації даних від внеску самого ансамблювання детекторів; кількісного пояснення через кореляцію передбачень моделей пулу та перевірки досяжності знання сцени реальним класифікатором у проаналізованій літературі також не наведено. Саме цю прогалину закриває це дослідження.

Матеріали та методи дослідження. Сценозалежне ансамблеве виявлення об'єктів на аерознімках з безпілотних літальних апаратів (БПЛА) є передовим підходом у комп'ютерному зорі, який адаптує комбінації різних моделей глибокого навчання (ансамблів) до специфічних умов зйомки (освітлення, висота, тип ландшафту, щільність об'єктів). Дослідження виконано на двох наборах даних аерознімків: набір даних VisDrone-DET 2019 [22] з 8-класовою розміткою та евристичним розподілом оцінювальних зображень за сценами (день – 441, ніч – 542, дощ – 91); набір даних UAVDT (англ. *Unmanned Aerial Vehicle Benchmark*) [5] з офіційними авторськими мітками умов (29 денних, 13 нічних, 8 туманних послідовностей; 1372 кадри), на який пул моделей перенесено без донавчання. Використано пул із семи донавчених на наборі даних VisDrone детекторів п'яти поколінь YOLO (25,9-68,2 млн параметрів) з відкритого модельного зоопарку у роботі [15]; передбачення зливалися методом WBF [17], якість вимірювалася метрикою mAP@50 у стилі COCO [11] за єдиним протоколом (поріг оцінювання детекторів 0,001, роздільність 1280), а статистичну значущість різниць оцінено пар-

ним та кластерним бутстрепом із крос-фітованим відбором конфігурацій.

Результати дослідження та їх обговорення / Research results and their discussion

Сценозалежне ансамблеве виявлення об'єктів (англ. *Scene-Conditioned Ensemble Object Detection*) на аерознімках з БПЛА – це передовий підхід у комп'ютерному зорі. Він адаптує роботу нейромереж під конкретні умови знімання (тип сцени, освітлення, щільність забудови). Концепція маршрутизації кадрів за сценами (англ. *Scene-based Frame Routing*) передбачає, що вхідний кадр спочатку класифікується за типом сцени, а потім направляється до найефективнішої для цієї сцени моделі (або комбінації моделей) з ансамблю. Проте цей підхід має чіткі межі корисності.

Постановка завдання та конвеєр. Нехай $M = \{m_1, \dots, m_k\}$ – пул детекторів, а S – простір сцен. Конфігурація ансамблю $c = (M', w)$ визначається підмножиною членів $M' \in M$ та вектором ваг злиття w ; результатом інференсу конфігурації зображення на кадрі $x \in$ множині рамок, отримана WBF-злиттям передбачень членів. Сценозалежна (маршрутизована) система задається вентиляльною функцією $g: X \rightarrow S$ (сценозалежним селектором, англ. *gating network*) та відображенням сцен у конфігурації зображення $s \in c(s)$: кадр x обробляється конфігурацією $c(g(x))$. Динамічне зважування довіри реалізується пер-сеновим вектором ваг w у $c(s)$: модель, що краща за умов сцени s , отримує більшу вагу під час злиття передбачень; граничним випадком м'якого зважування без відбору є злиття повного пулу. Агностична альтернатива використовує одну фіксовану конфігурацію c^* для всіх кадрів. Критерій порівняння архітектур – mAP@50 обох систем на одному пулі M за однакового протоколу; маршрутизація кадру вважається корисною, якщо її перевага статистично значуща і не відтворюється жодною фіксованою конфігурацією, зокрема – злиттям повного пулу. Загальну схему конвеєра та порівнюваної альтернативи наведено на рис. 1.

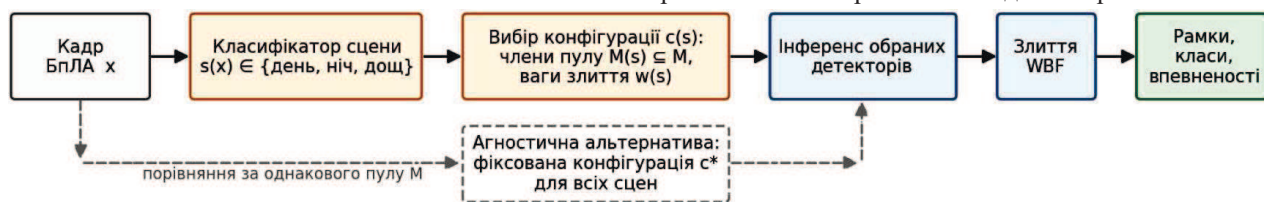


Рис. 1. Конвеєр сценозалежного ансамблевого виявлення та агностична альтернатива, з якою він порівнюється / Pipeline of scene-dependent ensemble detection and the scene-agnostic alternative it is compared against

Простір сцен не є довільним вибором авторів: на наборі даних VisDrone класи день/ніч/дощ задано евристичним розбиттям за ознаками яскравості, насиченості та контрасту, що відтворює домінуючі умови знімання набору (для ночі використано тестове розбиття, оскільки валідаційне містить тільки 16 нічних зображень), а на наборі даних UAVDT використано офіційні авторські атрибути умов (день/ніч/туман), які постають разом із набором. Потрібно чесно визнати, що клас "дощ" на наборі даних VisDrone представлений тільки 91 оцінювальним зображенням, тож пер-сенові оцінки для нього мають найширші довірчі інтервали. Класи *tricycle* та *awning-tricycle* вилучено з розмітки з огляду на малу представленість; сумісність розмітки кожної моделі з 8-класовою схемою перевіряється програмно перед інференсом.

Попередньо виявлено та виправлено три систематичні дефекти протоколу оцінювання детекторів. По-перше, mAP помилково обчислювався за порогом впевненості розгортання системи 0,25, що обрізає криву "точність – повнота" і сильніше штрафувало малі моделі: для YOLOv8n виправлення порогу оцінювання детекторів на 0,001 підвищило mAP на 42 %, тоді як для великих моделей ефект близький до нуля, тобто дефект спотворює саме порівняння моделей. По-друге, донавчені ваги видають 10 класів вихідної розмітки Набору даних VisDrone, і без перепризначення індексів класи автобусів і мотоциклів оцінювалися за хибним індексом (їхня покласова AP після виправлення зросла з майже нуля до 0,29-0,80). По-третє, вбудований шлях оцінювання детекторів `Ultralytics model.val()` має внутрішній ліміт часу на постоброблення (2,8 с/зображення), який

на щільних кадрах Набору даних VisDrone мовчки об'являє mAP великих моделей (для YOLOv8l – до 0,057 замість $\approx 0,55$); тому все оцінювання детекторів уніфіковано на шляху predict із власною реалізацією обчислення метрики.

Єдиний протокол усіх експериментів: поріг оцінювання детекторів 0,001, роздільність інференсу 1280, злиття передбачень – WBF з порогом IoU 0,55 і порогом відкидання 0,01; ваги членів однакові, якщо не зазначено інше. Швидкість (FPS) вимірюється окремо на роздільності розгортання системи 640 з порогом 0,25 цим самим вимірювачем для одиночних моделей та ансамблів. Власну реалізацію метрики звірено з еталонною русоtools на ідентичних передбаченнях (розбіжність не перевищує 0,0006 за mAP@50 та 0,0016 за mAP@50-95). Інференс кожної моделі виконується один раз на пару "модель-сцена" з кешуванням передбачень; злиття, підбір ваг і статистичні процедури працюють над кешем, що гарантує ідентичність входів для всіх порівнюваних конфігурацій.

Статистичну значущість різниць оцінено парним бутстрепом за зображеннями (1000 реплік): зіставлення передбачень з еталоном виконується один раз на зображення, репліки тільки переагрегують отримані прапорці, і всі порівнювані конфігурації зображення оці-

нюються на спільному ресемплі, тож різниці є парними. На наборі даних UAVDT, де кадри одного відео сильно корельовані, застосовано кластерний бутстреп із ресемплінгом на рівні послідовностей; кадри відео проріджено до одного кадру на секунду, передбачення пулу зведено до класів набору даних UAVDT (car, truck, bus; фургони злито з car відповідно до визначення категорії car в набір даних UAVDT), а рамки в офіційних ignore-зонах відкинуто за центром. Для конфігурацій із відбором членів за даними використано крос-фітування: конфігурація ансамблю для послідовності обирається без використання її кадрів, що усуває оптимістичний зсув відбору. Досяжність знання сцени оцінено незалежним CNN-класифікатором сцен MobileNetV3-Small (латентність близько 9 мс/кадр) – легка згорткова нейронна мережа, що відіграє роль класифікатора контексту сцени (англ. Scene Context Classifier) і перед виявленням об'єктів аналізує кадр загалом, визначаючи умови знімання: на наборі даних UAVDT – проти офіційних міток умов п'ятифолдовою крос-валідацією за послідовностями, що унеможлиблює витік через майже-дублікації сусідніх кадрів.

Програму експериментів наведено в табл. 1, де кожен наступний експеримент знімає одне обмеження попереднього.

Табл. 1. Програма експериментів / The program of experiments

Код	Експеримент	Що перевіряє
E0	Пер-сценова оцінка одиночних моделей	Базова лінія
E1	Пер-сценовий ансамбль детекторів (оракульна маршрутизація кадру)	Чи допомагає ансамбль детекторів
E3	Сценонезалежний ансамбль детекторів, однакові ваги	Чи важлива маршрутизація кадру
E4	Сценонезалежний ансамбль детекторів, підібрані ваги	Чи дає щось підбір ваг
E5	Крос-архітектур детекторівний ансамбль детекторів (+DETR)	Чи допомагає різноманітність
E6	Маршрутизація кадру: оракул проти CNN-класифікатора	Вартість помилок Класифікатора
E7a	Пер-сценова оцінка гетерогенного пулу (4 покоління YOLO)	Чи існує сильніший і різноманітніший пул
E7b	Маршрутизація кадру на гетерогенному пулі	Чи виграє маршрутизація кадру, коли сцени обирають різні моделі
E7c	Гейт крос-парадигмальних кандидатів (RT-DETR / DETA / DETR)	Чи існують публічні сильні не-YOLO ваги
E7d	Парний бутстреп (1000 реплік)	Статистична значущість ключових різниць
E8	Нульове перенесення пулу на наборі даних UAVDT (офіційні мітки умов)	Відтворюваність без циклічності розбиття; бюджетна ніша
E8b	Класифікатор умов на наборі даних UAVDT (проти офіційних міток)	Чи розгортається бюджетна ніша реальним класифікатором

Базова лінія: одиночні моделі. Якість одиночних моделей YOLOv8 на повному валідаційному наборі даних VisDrone наведено в табл. 2: вона монотонно зростає з розміром моделі, насичуючись після YOLOv8m. Найкраща одиночна модель за сценами: день – YOLOv8l (0,544), ніч – YOLOv8l (0,334), дощ – YOLOv8x (0,597); саме нерівномірність цих значень і мотивує сценозалежний підхід.

Табл. 2. Одиночні моделі детектора YOLOv8 на валідаційному наборі даних VisDrone (роздільність 1280) / Single YOLOv8models on the VisDrone validation set (resolution 1280)

Модель детектора	Параметри, млн	mAP@50	mAP@50-95
YOLOv8n	3,2	0,431	0,257
YOLOv8s	11,2	0,492	0,298
YOLOv8m	25,9	0,530	0,325
YOLOv8l	43,7	0,547	0,338
YOLOv8x	68,2	0,551	0,342

Ансамблювання детекторів проти одиночної моделі. Ансамблювання детекторів (Ensembling) зазвичай перевершує одиночну модель за точністю та стабільністю, оскільки об'єднує прогнози кількох різних алгорит-

мів для зменшення помилок (зсуву та дисперсії). Проте одиночна модель значно вирає у швидкості роботи та простоті впровадження.

Табл. 3. Головні конфігурації базового пулу на наборі даних VisDrone (mAP@50; "сер." – середнє, зважене за кількістю зображень) / Main configurations of the base pool on VisDrone (mAP@50; "avg." – image-weighted mean)

Конфігурація ансамблю детекторів	День	Ніч	Дощ	Сер.
Одиночна найкраща (на сцену)	0,544	0,334	0,597	0,443
Ансамбль детекторів l/x/m, агностичний, однакові ваги	0,600	0,382	0,649	0,494
Ансамбль детекторів l/x/m, агностичний, підібрані ваги	0,599	0,381	0,651	0,493
Ансамбль детекторів l/x/m, адаптивний (оракул)	0,600	0,381	0,651	0,494
Ансамбль детекторів l/x/m + DETR	0,600	0,380	0,650	0,493

Основні методи об'єднання в ансамблях: Non-Maximum Suppression (NMS) – залишає тільки рамку з найвищою впевненістю, видаляючи інші, що сильно перекриваються; Weighted Boxes Fusion (WBF) – новий

підхід, який не просто відкидає "слабші" рамки, а усереднює їхні координати та коефіцієнти впевненості, що дає точніший результат.

Ансамбль детекторів WBF трьох найбільших моделей YOLOv8 (l/x/m) з однаковими вагами перевершує найкращу одиночну модель у кожній сцені (табл. 3): зважене за кількістю зображень середнє зростає від 0,443 до 0,494 (+0,051). Ціна – приблизно триразове зниження швидкості (5,7 FPS проти 16,5 FPS у найкращій одиночній моделі вдень), тобто це компроміс Парето

то: обидві точки не доміновані, і вибір між ними залежить від обчислювального бюджету (рис. 2). Підбір ваг членів на валідаційній підмножині не дає нічого понад однакові ваги (0,493 проти 0,494), а оракульна пер-сцена маршрутизація кадру – нічого понад глобальний ансамбль детекторів (0,494 проти 0,494). Додавання трансформера DETR також не змінює результату ($\pm 0,001$): пошук ваг сам присвоює DETR мінімальну вагу, оскільки модель, обмежена 100 запитами на кадр, надто слабка для щільних сцен (0,055–0,115 mAP@50 за сценами).

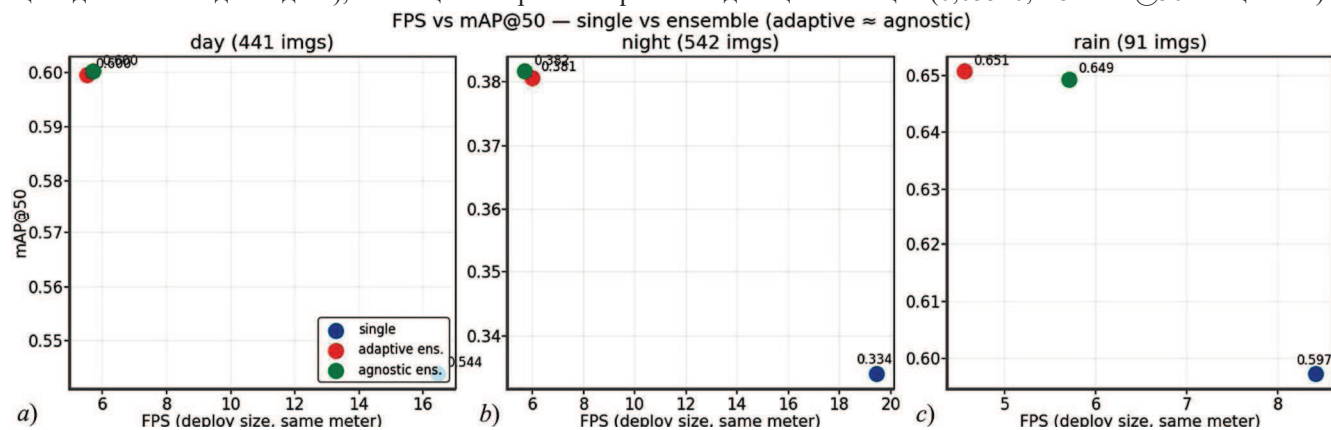


Рис. 2. Компроміс Парето "якість-швидкість": одиночні моделі детектора YOLOv8 та ансамбль детекторів WBF (набір даних VisDrone) / The quality-speed Pareto trade-off: single YOLOv8 models vs the WBF ensemble (VisDrone dataset)

Покласовий аналіз та розмір об'єктів. Ансамбль детекторів перевершує одиночну модель на кожному з восьми класів (рис. 3), до того ж приріст концентрується на складних класах транспорту: автобус +0,082, вантажівка +0,076, фургон +0,073, мотоцикл +0,063; на найпростішому класі (легковий автомобіль) приріст найменший (+0,029). Важливий нюанс виявлено на порозі розгортання системи ($\text{conf} \geq 0,25$): повнота виявлення дрібних об'єктів (менших за 32×32 пікселі) в ансамблі трохи знижується (0,590 $\rightarrow$ 0,573), тоді як для середніх і великих зростає (+0,015 і +0,029 відповідно).

Причина в тому, що WBF усереднює впевненість збіжних рамок, тож дрібний об'єкт, знайдений тільки однією моделлю з трьох, опускається нижче від робочого порогу. Виграш ансамблю реальний на рівні ранжування передбачень, але для дрібних об'єктів на робочому порозі він не безкоштовний. Якісні приклади детекції за чотирьох умов знімання наведено на рис. 4: візуально передбачення одиночної моделі та ансамблю близькі, що узгоджується з кількісними різницями; систематичні відмінності видно на середніх і великих об'єктах та в повноті покриття щільних груп.

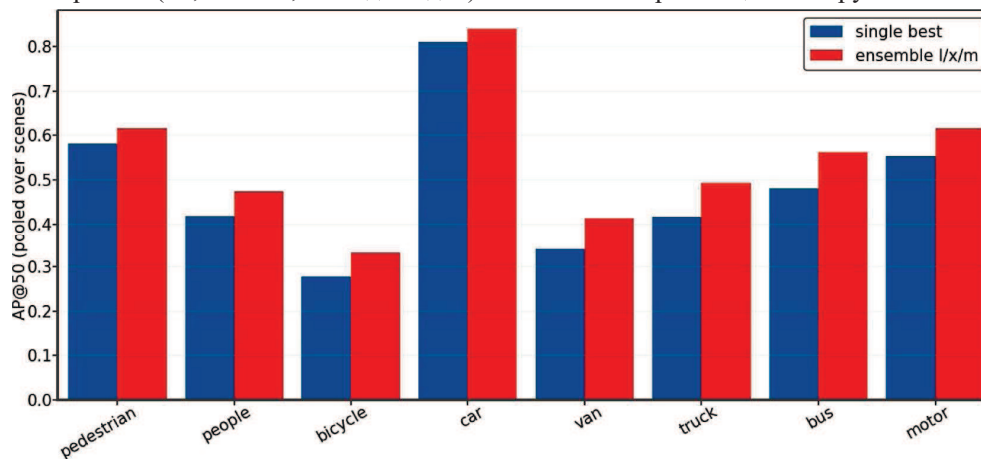


Рис. 3. Покласова AP@50: найкраща одиночна модель проти ансамблю WBF (набір даних VisDrone) / Per-class AP@50: the best single model vs the WBF ensemble (VisDrone dataset)

Гетерогенний пул: чотири покоління детекторів YOLO. Чотири покоління детекторів YOLO (англ. *You Only Look Once*) відображають еволюцію цієї архітектури комп'ютерного зору від першої революційної ідеї до сучасних його стандартів. Моделі YOLO змінили підхід до виявлення різного роду об'єктів (Object Detection), відмовившись від двоетапних алгоритмів (як R-CNN) на користь їх розпізнавання за один прохід (Single-Shot), що забезпечило роботу комп'ютерних систем у реальному часі.

Головне заперечення до негативного результату маршрутизації даних на базовому пулі – що надлишковість могла бути артефактом одного сімейства (усі члени – YOLOv8). Тому аналіз відтворено на пулі з чотирьох новіших поколінь YOLO з модельного зоопарку у роботі [15], тренуваних за єдиним рецептом (300 епох, роздільність 640), тож відмінності між ними зумовлені архітектурою детекторів, а не режимом тренування (табл. 4). Кожна модель нового покоління перевершує кожну модель YOLOv8 у кожній сцені, до того ж впер-

ше різні сцени віддають перевагу різним сильним моделям: удень найкращою є YOLO26x (0,583), уночі та в дощ – YOLOv9e (0,358 і 0,624). Новий найкращий агностичний ансамбль детекторів – рівноваговий WBF

{YOLOv9e, YOLO26x, YOLOv10x}: день 0,645, ніч 0,421, дощ 0,684, зважене середнє 0,535 проти 0,494 у l/x/m (+0,041, 95 % ДІ [+0,035; +0,047], значуще в кожній сцені окремо).

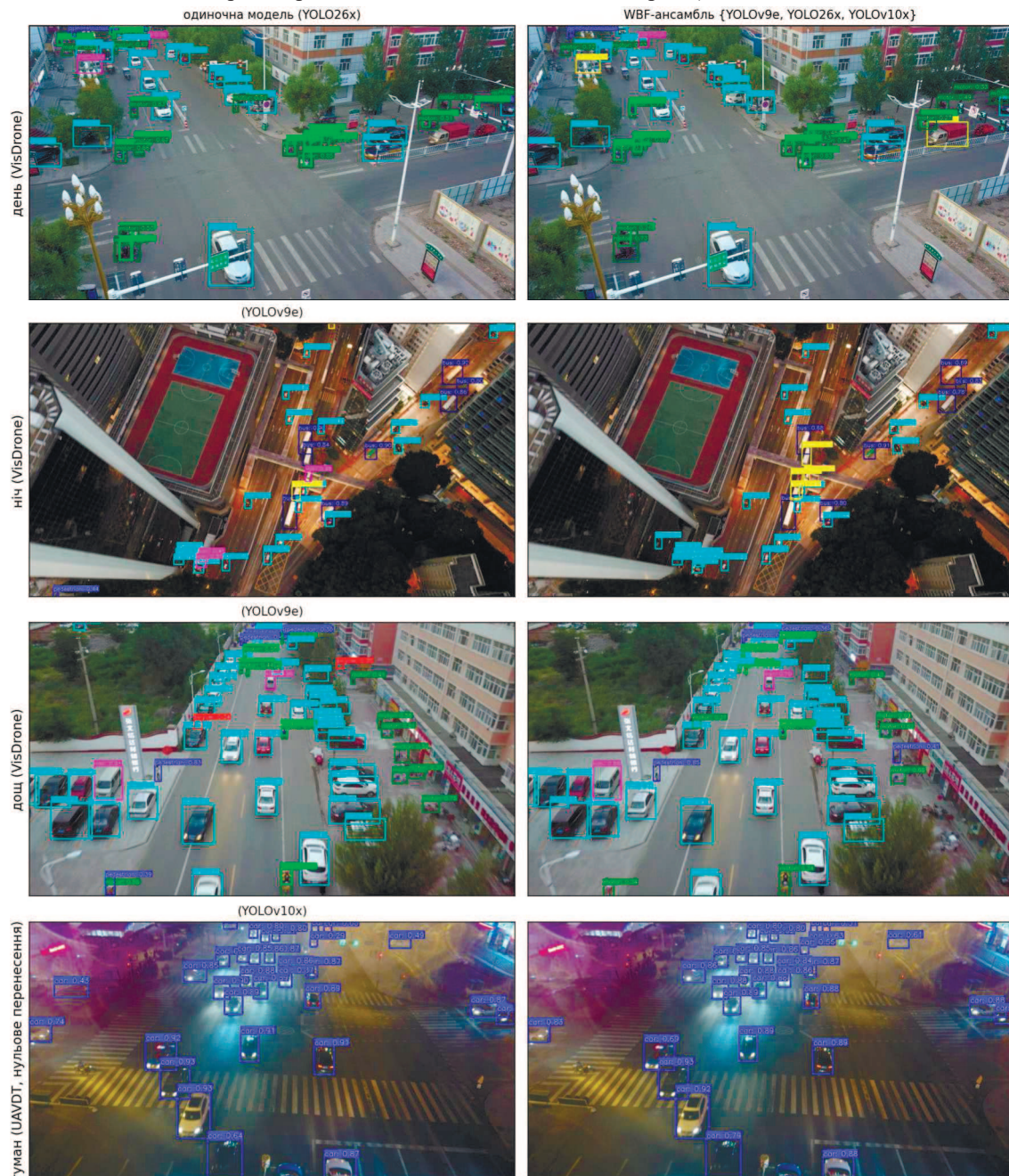


Рис. 4. Приклади детекції ($\text{conf} \geq 0,25$): найкраща одиночна модель сцени проти WBF-ансамблю. Нижній рядок – туман (набір даних UAVDT): умова, яку класифікатор сцен найчастіше плутає з днем (76 зі 175 кадрів) / Detection examples ($\text{conf} \geq 0.25$): the best single model of each scene vs the WBF ensemble; bottom row – fog (UAVDT), the condition most often confused with day by the scene classifier (76 of 175 frames)

Табл. 4. Гетерогенний пул: $\text{mAP}@50$ одиночних моделей за сценами (набір даних VisDrone) / Heterogeneous pool: per-scene $\text{mAP}@50$ of single models (VisDrone dataset)

Модель детекторів	Параметри, млн	День	Ніч	Дощ	FPS@640 (день)
YOLOv8m	25,9	0,517	0,309	0,571	29,3
YOLOv8l	43,7	0,537	0,326	0,566	16,8
YOLOv8x	68,2	0,535	0,323	0,592	14,0
YOLOv9e	58,2	0,567	0,358	0,624	10,3
YOLOv10x	31,7	0,573	0,351	0,612	16,2
YOLO11x	56,9	0,561	0,346	0,593	12,7
YOLO26x	58,8	0,583	0,351	0,612	13,0

Гетерогенний пул: чотири покоління детекторів YOLO. Головне заперечення до негативного результату маршрутизації даних на базовому пулі – що надлишковість могла бути артефактом одного сімейства (усі члени – YOLOv8). Тому аналіз відтворено на пулі з чотирьох новіших поколінь YOLO з модельного зоопарку у роботі [15], тренуваних за єдиним рецептом (300 епох, роздільність 640), тож відмінності між ними зумовлені архітектурою детекторів, а не режимом тренування (табл. 4). Кожна модель нового покоління перевершує кожну модель YOLOv8 у кожній сцені, до того ж вперше різні сцени віддають перевагу різним силь-

ним моделям: удень найкращою є YOLO26x (0,583), уночі та в дощ – YOLOv9e (0,358 і 0,624). Новий найкращий агностичний ансамбль детекторів – рівноваговий WBF {YOLOv9e, YOLO26x, YOLOv10x}: день 0,645, ніч 0,421, дощ 0,684, зважене середнє 0,535 проти 0,494 у l/x/m (+0,041, 95 % ДІ [+0,035; +0,047], значуще в кожній сцені окремо).

Табл. 4. Гетерогенний пул: mAP@50 одиночних моделей за сценами (набір даних VisDrone) / Heterogeneous pool: per-scene mAP@50 of single models (VisDrone dataset)

Модель детекторів	Параметри, млн	День	Ніч	Дощ	FPS@640 (день)
YOLOv8m	25,9	0,517	0,309	0,571	29,3
YOLOv8l	43,7	0,537	0,326	0,566	16,8
YOLOv8x	68,2	0,535	0,323	0,592	14,0
YOLOv9e	58,2	0,567	0,358	0,624	10,3
YOLOv10x	31,7	0,573	0,351	0,612	16,2
YOLO11x	56,9	0,561	0,346	0,593	12,7
YOLO26x	58,8	0,583	0,351	0,612	13,0

Маршрутизація кадру не перевершує глобального ансамблю детекторів. Головний результат наведено в табл. 5. Маршрутизована одиночна модель значуще перевершує глобально найкращу одиночну (+0,006, ДІ [+0,002; +0,011]) – сценоспецифічна перевага існує. Однак для ансамблів перевага маршрутизації даних так і не з'являється: оракульна маршрутизація кадру (ідеальне знання сцени, пер-сеновий підбір членів і ваг) дає -0,0006 mAP@50 (ДІ [-0,0011; -0,0002]) відносно агностичного ансамблю, до того ж вдень вона значуще гірша (-0,0016, ДІ [-0,0025; -0,0006]) – пер-сеновий підбір ваг перенавчається на малій підмножині. Пул-рівневий підбір ваг знову вироджується в однакові ва-

ги. Механізм видно точно: сценоспецифічна перевага існує, але злиття повного пулу відновлює її повністю і зі запасом (+0,062 над маршрутизованою одиночною, ДІ [+0,057; +0,068]) – ансамблювання детекторів "поглинає" маршрутизацію.

Табл. 5. Головні конфігурації гетерогенного пулу на наборі даних VisDrone (mAP@50; "сер." – середнє, зважене за кількістю зображень) / Main configurations of the heterogeneous pool on VisDrone (mAP@50; "avg." – image-weighted mean)

Конфігурація пулу	День	Ніч	Дощ	Сер.
Одиночна, глобально найкраща (YOLOv9e)	0,567	0,358	0,624	0,466
Одиночна, маршрутизована (день → YOLO26x)	0,583	0,358	0,624	0,473
Ансамбль детекторів {9e, 26x, 10x}, агностичний	0,645	0,421	0,684	0,535
Ансамбль детекторів, маршрутизований (оракул)	0,644	0,421	0,684	0,534
Довідково: ансамбль детекторів YOLOv8l/x/m	0,600	0,382	0,649	0,494

Механізм: кореляція передбачень. Пояснення дає попарне перекриття передбачень – частка рамок моделі, що мають збіг того самого класу з IoU $\geq 0,5$ у передбаченнях іншої моделі (рис. 5, ліва панель). Для всіх пар пулу воно знаходиться у вузькій смузі 0,78-0,81: між поколіннями детекторів YOLO так само, як і всередині сімейства детекторів YOLOv8, і навіть між різними рецептами донавчання. Отже, у межах перевіреного простору донавчених ваг надлишковість пулу є властивістю навчального розподілу даних, а не архітектури детектора чи рецепта тренування. DETR справді декорельований (перекриття $\approx 0,28$ з кожною моделлю пулу), але заслабкий, щоб цим скористатися.

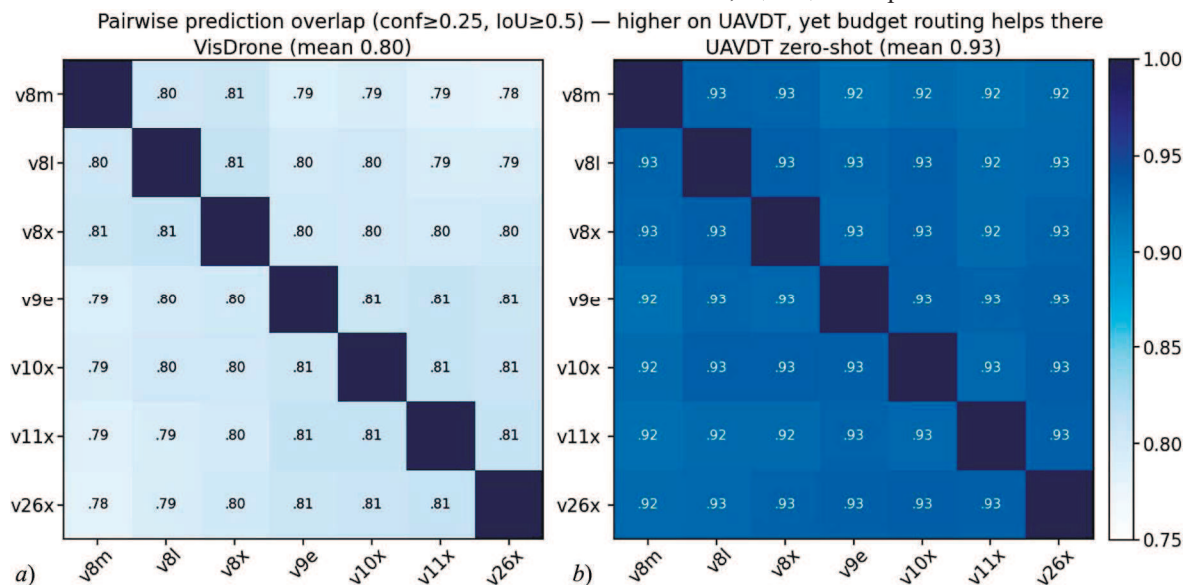


Рис. 5. Попарне перекриття передбачень моделей пулу: набір даних VisDrone (ліворуч, середнє 0,80) та набір даних UAVDT (праворуч, середнє 0,93) / Pairwise prediction overlap of the pool models: VisDrone (left, mean 0.80) and UAVDT (right, mean 0.93)

Крос-парадигмальні кандидати (моделі) комп'ютерного зору – це передові архітектури штучного інтелекту, які об'єднують різні фундаментальні парадигми обробки інформації. Найчастіше йдеться про поєднання згорткових нейромереж (CNN), які мають сильне локальне індуктивне зміщення, та трансформерів (ViT), здатних моделювати глобальні залежності через механізми самоуваги. Також до цієї категорії відносять моделі, що поєднують нейросимволічний підхід (програмування/логіка) із глибоким навчанням або узгоджують

просторово-канальне кодування із глобально-локальним сприйняттям.

Останнє незакрите питання – чи існує трансформерний детектор, водночас сильний і декорельований. Вичерпний гейт усіх публічних трансформерних ваг набору даних VisDrone (табл. 6) показує, що жоден кандидат не проходить: репозиторії або порожні, або містять дегенеративні чи іграшкові моделі. Попри сильні трансформерні результати в літературі [20], клітинка "сильний і крос-парадигмальний" у просторі публічно дос-

тупних ваг станом на липень 2026 р. порожня, тож умову "сильний і декорельований" неможливо перевірити готовими вагами – тільки власним донавчанням.

Табл. 6. Гейт публічних трансформерних ваг набору даних VisDrone (станом на липень 2026 р.) / Gate of public transformer weights for VisDrone (as of July 2026)

Кандидат	Стан	mAP@50 (день/ніч/дощ)	Перекриття
RT-DETR-L (dgural)	Порожній репозиторій	–	–
RT-DETR (silveroupti)	Порожній репозиторій	–	–
DETA-Swin-L (dani880)	Дегенеративний	0,009 / 0,010 / 0,008	≈0,01
DETR (sameerrind009)	Іграшковий	0,034 / 0,020 / 0,045	≈0,11
DETR (okhytrov)	Заслабкий	0,106 / 0,055 / 0,115	≈0,28

Табл. 7. Парний бутстреп за зображеннями, набір даних VisDrone (1000 реплік) / Paired bootstrap over images, VisDrone dataset (1000 replicates)

Різниця (mAP@50)	Сер. сцен [95 % ДІ]	Зважено [95 % ДІ]	P ($\Delta \leq 0$)
Новий пул – 1/x/m	+0,0397 [+0,0323; +0,0467]	+0,0412 [+0,0350; +0,0472]	0,00
Маршрутизація кадру – агностичний	–0,0005 [–0,0011; +0,0001]	–0,0006 [–0,0011; –0,0002]	0,92
Маршрутизація кадру одиночна – глобальна одиночна	+0,0049 [+0,0015; +0,0086]	+0,0061 [+0,0018; +0,0106]	0,00
Ансамбль детекторів – маршрутизація одиночна	+0,0620 [+0,0553; +0,0684]	+0,0626 [+0,0568; +0,0680]	0,00

Перенесення на наборі даних UAVDT. Нульове перенесення пулу на наборі даних UAVDT з офіційними мітками умов дало якісно нову картину (табл. 8): уподобання сцен уперше різко розходяться. Детектор YOLO11x виявляється "денним спеціалістом" (0,642 удень проти 0,426 у тумані), а ніч і туман обирають детектор YOLOv10x. Глобальний топ-3 – знову {YOLOv9e, YOLOv10x, YOLO26x}, як і на наборі даних VisDrone.

Табл. 8. Набір даних UAVDT, нульове перенесення: mAP@50 одиночних моделей за офіційними мітками умов / UAVDT dataset, zero-shot transfer: mAP@50 of single models by official condition labels

Модель детектора	День	Ніч	Туман
YOLOv8l	0,578	0,580	0,438
YOLOv8x	0,586	0,555	0,435
YOLOv8m	0,632	0,544	0,497
YOLOv9e	0,647	0,575	0,507
YOLOv10x	0,610	0,619	0,558
YOLO11x	0,642	0,570	0,426
YOLO26x	0,620	0,580	0,521

Ключові конфігурації моделі наведено в табл. 9, кластерний бутстреп – у табл. 10. За чесного крос-фітованого відбору (конфігурація для послідовності оби-

Табл. 9. Конфігурації моделі на наборі даних UAVDT, нульове перенесення (mAP@50; "сер. (зваж.)" – зважене за кількістю кадрів) / Model configurations on UAVDT, zero-shot transfer (mAP@50; "avg. (weighted)" – frame-weighted mean)

Конфігурація моделі	День	Ніч	Туман	Сер. (зваж.)
Одиночна, глобально найкраща (YOLOv9e)	0,649	0,573	0,502	0,609
Одиночна, маршрутизована (крос-фіт)	0,637	0,552	0,541	0,601
Ансамбль детекторів топ-3, агностичний {9e, 10x, 26x}	0,646	0,609	0,579	0,627
Ансамбль детекторів топ-3, агностичний (крос-фіт)	0,649	0,579	0,544	0,616
Ансамбль детекторів топ-3, маршрутизований (крос-фіт)	0,648	0,590	0,585	0,624
Ансамбль детекторів з усіх 7 моделей, без відбору	0,661	0,610	0,567	0,634

Табл. 10. Кластерний бутстреп на рівні послідовностей, набір даних UAVDT (1000 реплік) / Cluster bootstrap at the sequence level, UAVDT dataset (1000 replicates)

Різниця (mAP@50)	Сер. сцен [95 % ДІ]	Зважено [95 % ДІ]
Маршрутизація даних топ-3 – агност. топ-3 (обидва крос-фіт)	+0,0141 [+0,0007; +0,0269]	+0,0081 [–0,0043; +0,0218]
Маршрутизація даних топ-3 (крос-фіт) – усі-7	–0,0071 [–0,0218; +0,0054]	–0,0105 [–0,0244; +0,0026]
Усі-7 – агност. топ-3	+0,0061 [–0,0037; +0,0183]	+0,0097 [+0,0001; +0,0204]
Маршрутизація даних одиночна (крос-фіт) – глобальна одиночна	+0,0003 [–0,0303; +0,0223]	–0,0105 [–0,0369; +0,0062]

Статистичні гарантії. Ключові різниці малі, тож точкових оцінок недостатньо; підсумок парного бутстрепа наведено в табл. 7. Тут парний бутстреп за зображеннями (англ. *Paired Bootstrap over Images*) – це статистичний метод повторної вибірки (resampling), який використовується для оцінювання невизначеності, обчислення довірчих інтервалів або статистичного порівняння якості моделей комп'ютерного зору (наприклад, YOLO, Faster R-CNN) на фіксованому тестовому або валідаційному наборі даних.

З таблиці видно, що приріст гетерогенного пулу та приріст ансамблювання детекторів значущі з великим запасом; перевага маршрутизованої одиночної моделі значуща, але мала; різниця "маршрутизація кадру – агностичний ансамбль детекторів" значуща не відрізняється від нуля на збалансованому середньому і значуща від'ємна на зваженому.

рається без використання її кадрів) маршрутизований топ-3 ансамбль детекторів значуще перевершує агностичний топ-3 на збалансованому за сценами середньому (+0,0141, ДІ [+0,0007; +0,0269]) з найбільшим вигравом у тумані – це перший статистично значущий позитивний результат маршрутизації даних в роботі. Проте злиття всіх семи моделей без жодного відбору знову поглинає цю перевагу (–0,0071 відносно маршрутизованого топ-3, незначуще). Повчальний урок про зсув відбору: "оракульна" перевага маршрутизованої одиночної моделі (+0,022) після крос-фітування зникає повністю.

Оскільки специфікація ракурсів БПЛА (маленькі об'єкти, велика щільність, зміна кутів огляду та рух камери) суттєво відрізняється від класичних наземних ракурсів, правильне налаштування процесу перенесення є критично важливим. Показово, що попарне перекриття передбачень на наборі даних UAVDT становить 0,92–0,93 (див. рис. 5, права панель) – вище, ніж 0,78–0,81 на наборі даних VisDrone, – і все ж маршрутизація кадру за фіксованого бюджету допомагає (рис. 6). Отже, вирішальним є не рівень надлишковості сам собою, а те, чи переживають сценозалежні відмінності рангу моделей злиття повного пулу.

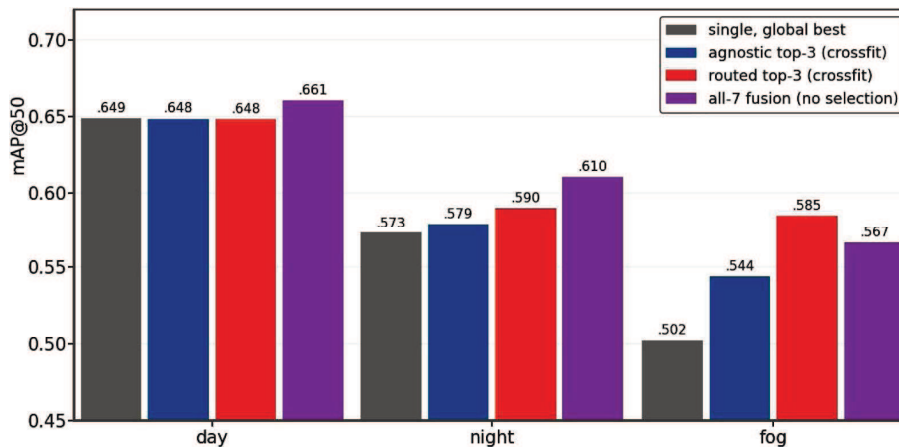


Рис. 6. Набір даних UAVDT: за фіксованого бюджету top-3 маршрутизація кадру допомагає (найбільше в тумані); злиття всього пулу поглинає перевагу / UAVDT dataset: at a fixed top-3 budget frame routing helps (most in fog); fusing the whole pool absorbs the advantage

Реальний класифікатор сцен (класифікатор контексту сцени). Останнє питання – чи досягне знання сцени, якого потребує маршрутизація кадру. На наборі даних VisDrone незалежний CNN-класифікатор MobileNetV3-Small досягає узгодженості з мітками розбиття 88,7 % (день 92,3 %, ніч 92,6 %, але дощ – тільки 48,4 %; матриця сплутування – рис. 7) за латентності 8,9 мс/зображення. Для ансамблів вартість його помилок практично нульова ($-0,0001$, табл. 11): помилково маршрутизоване зображення отримує майже ідентичний ансамбль детекторів. Для одиночних моделей помилки коштують $-0,0038$, і оракульна перевага маршрутизованої одиночної моделі на середньому за сценами стає незначущою. Виправлене тренування моделі, узгоджене з простором міток (без колірної аугментації та синтетики), підвищує точність на дощі до 54,9 % і, головне, переміщує помилки туди, де вони безкоштовні (ніч і дощ маршрутизуються до тієї самої моделі): вартість падає до $+0,0001$, а перевага відновлюється до $+0,0050$ ($P(\Delta \leq 0) = 0,054$).

Табл. 11. Маршрутизація кадру реальним CNN-класифікатором на наборі даних VisDrone (mAP@50; "сер." – середнє сцен) / Frame routing by a real CNN classifier on VisDrone (mAP@50; "avg." – scene mean)

Конфігурація	День	Ніч	Дощ	Сер.
Ансамбль детекторів, агностичний	0,6452	0,4205	0,6841	0,5833
Ансамбль детекторів, маршрутизований (оракул)	0,6437	0,4205	0,6841	0,5828
Ансамбль детекторів, маршрутизований (CNN)	0,6438	0,4203	0,6841	0,5827
Одиночна, маршрутизована (оракул)	0,5829	0,3580	0,6239	0,5216
Одиночна, маршрутизована (CNN)	0,5798	0,3580	0,6152	0,5177

Вирішальний тест виконано на відомому наборі даних UAVDT проти офіційних міток умов (п'ятифолдова крос-валідація за послідовностями): класифікатор розпізнає день на 99,8 % і ніч на 96,1 %, але туман – тільки на 32,0 % покaдрово. Оскільки це справжні авторські мітки, низька класифікованість туману є внутрішньою складністю, а не шумом псевдоміток: туман – атмосферна вуаль, ортогональна осі освітленості, і його плутають і з днем, і з ніччю (приклад туманного кадру – нижній рядок на рис. 4). Ціна руйнівна саме там, де живе бюджетна ніша: помилково маршрутизовані туманні кадри отримують денну конфігурацію, що містить

найгіршу для туману модель (YOLO11x, 0,426), і mAP туману падає від 0,579 до 0,526. Розгортувана маршрутизація кадру проти агностичного top-3 дає $-0,0004$ замість $+0,0115$ за офіційними мітками – ніша зникає. Практичний шлях розгортання системи для маршрутизації даних – екзогенні сигнали умов (час доби, метадані погоди) замість класифікації зображення.

Scene classifier confusion (VisDrone)
overall 88.7%, 8.9 ms/img

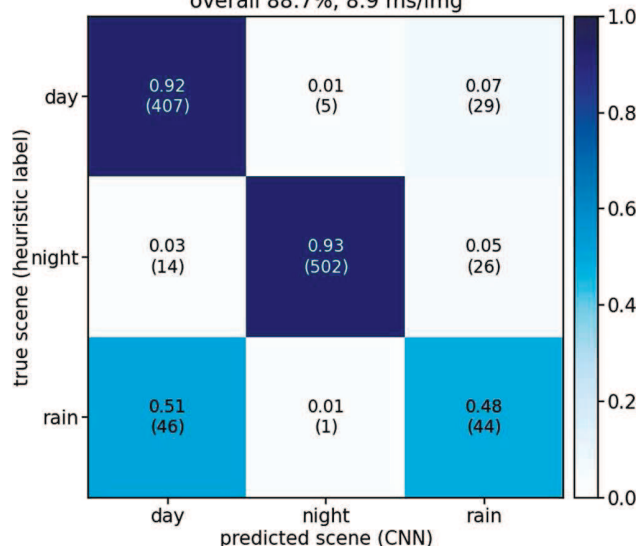


Рис. 7. Матриця сплутування CNN-класифікатора сцен на наборі даних VisDrone: клас "дощ" плутається з "днем" / Confusion matrix of the CNN scene classifier on VisDrone: the "rain" class is confused with "day"

Сукупність експериментів складається в цілісну картину. Ансамблювання детекторів методом WBF дає великий, статистично значущий і відтворюваний на двох наборах даних виграв, але це виграв ансамблювання детекторів, а не сценозалежності: у жодній із перевірених конфігурацій обох пулів маршрутизація кадру не додала нічого понад глобальний ансамбль детекторів тих самих моделей. Уточнена умова корисності маршрутизації даних полягає в тому, що недостатньо, щоб різні сцени віддавали перевагу різним сильним моделям – сценоспецифічна перевага має бути невідновлюваною злиттям повного пулу. За наявності бюджету на повний ансамбль детекторів злиття "поглинає" маршрутизацію; практичною нішею маршрутизації даних залишається жорсткий обчислювальний бюджет за доменного зсуву, де вона значуще перевершує агностичний ансамбль детекторів того самого розміру.

Окремо оцінимо обчислювальну ціну маршрутизації даних. Сім моделей пулу мають 25,9-68,2 млн параметрів (сумарно близько 343 млн), а інференс ансамблю коштує приблизно втричі більше за одиночну модель (див. табл. 4). Проте сама маршрутизація кадру майже нічого не додає до вартості агностичного ансамблю: члени пулу ті самі й додаткового донавчання не потребують, пер-сценово підбираються тільки ваги WBF над закешованими передбаченнями, а єдиний додатковий компонент – класифікатор сцен MobileNetV3-Small з латентністю 8,9 мс/кадр. Саме тому негативний результат такий показовий: не окупається навіть дешева надбудова, яка не збільшує ані вартості навчання, ані помітно вартості інференсу. Типовий прикладний сценарій, де цей висновок має значення, – бортовий або крайовий (edge) інференс на обмеженому обчислювальному ресурсі за змінних умов знімання: час доби відомий з годинника системи, а стан погоди – з метеоданих чи телеметрії, тож екзогенна маршрутизація кадру не потребує жодної додаткової моделі.

Обмеження результатів дослідження. Перенесення на набір даних UAVDT є крос-доменним: моделі, донавчені безпосередньо в домені UAVDT, могли б поводитися інакше; туман представлений тільки 8 послідовностями; злиття класу фургонів із легковими автомобілями – задокументоване наближення. Точність CNN-класифікатора на наборі даних VisDrone вимірюється відносно евристичних міток розбиття, тож частина "помилки" відображає межову неоднозначність самих сцен. Висновки стосуються фіксованого зовнішнього вентиля над незалежно донавченими детекторами і не переносяться на навчувані наскрізно МоЕ-архітектури, а твердження про надлишковість і відсутність сильних декорельованих кандидатів обмежені простором публічних донавчених ваг, доступних станом на липень 2026 р. Напрями подальшої роботи: донавчання сильного крос-парадигмального детектора (наприклад, RF-DETR або UAV-DETR [20]) на наборі даних VisDrone; донавчання пулу в домені UAVDT для перевірки бюджетної ніші без доменного зсуву; м'яка ймовірнісна маршрутизація кадру замість жорсткої.

Обговорення результатів дослідження. Сценозалежне ансамблеве виявлення об'єктів (англ. *Scene-Dependent Ensemble Object Detection*) – це передовий підхід у комп'ютерному зорі, де для розпізнавання об'єктів використовують не одну нейромережу, а комбінацію (ансамбль детекторів) кількох моделей, яка адаптується до конкретного типу сцени чи оточення. Замість агрегації відповідей усіх моделей однаково для будь-якого кадру, система аналізує глобальний контекст зображення (наприклад: населений пункт, траса, ліс, офіс, ніч, дощ) і динамічно змінює вагу або вибір моделей в ансамблі.

У роботі [23] представлено детектор TPH-YOLOv5 із трансформерними головами передбачень для аерознімків і показано приріст близько 7 % над базовою моделлю на наборі даних VisDrone, до того ж фінальне рішення поєднує кілька моделей; отриманий там інтегральний вигравш узгоджується з нашим природом ансамблювання детекторів (+0,041...+0,051 mAP@50), однак внесок власне ансамблювання детекторів у цій роботі не відокремлено від внеску архітектур детекторних змін, тоді як у нашому дослідженні ці внески ізольовано зі статистичними гарантіями.

У звіті змагання [3] узагальнено результати VisDrone-DET2021 і зазначено, що більшість найсильніших рішень поєднують передбачення кількох детекторів; це незалежно підтверджує практичну домінантність ансамблювання детекторів на щільних аерознімках, яку ми відтворюємо, – проте жодне з рішень змагання не порівнює сценозалежну маршрутизацію з агностичним ансамблем того самого пулу, тобто питання, на яке відповідає наша робота, у роботі [3] навіть не порушується.

У дослідженні [1] запропоновано нарізання зображення з дообробленням (SAHI) і показано приріст 5-7 AP для трьох детекторів на наборі даних VisDrone без жодного донавчання; це ортогональна до ансамблювання детекторів вісь покращення, спрямована саме на дрібні об'єкти, і вона добре доповнює наше спостереження, що WBF-ансамбль детекторів на робочому порозі трохи знижує повноту виявлення дрібних об'єктів (0,590 → 0,573): для систем, критичних до дрібних об'єктів, поєднання нарізання з ансамблюванням детекторів є природним напрямом.

У роботі [12] розроблено детектор IA-YOLO з диференційовним модулем адаптивного попереднього оброблення зображення, який навчається спільно з детектором і дає вигравш у тумані та за слабкого освітлення; це приклад адаптивності, вбудованої в модель на рівні входу, і він не суперечить нашому результату – навпаки, узгоджується зі сформульованою умовою: перевага IA-YOLO є невідновлюваною злиттям фіксованих моделей, бо жоден член пулу не бачить покращеного зображення.

У роботі [6] систематизовано динамічні нейронні мережі, які адаптують структуру чи параметри до кожного входу, і зазначено, що вигравш умовних обчислень виникає тоді, коли для частини входів обчислення надлишкові; наш результат уточнює цю картину для адаптивності на рівні цілих детекторів: коли передбачення членів пулу надлишкові (перекриття 0,78–0,93), вибір моделі за сценою не додає якості виявлення понад злиття, і вигравш з'являється тільки за доменного зсуву з фіксованим бюджетом.

У дослідженні [10] запропоновано доменну адаптацію детектора з дня на ніч без розмітки нічних зображень (2PCNet) і показано істотний приріст на нічних даних; це альтернативний до маршрутизації даних шлях врахування умов – адаптувати одну модель, а не обирати серед фіксованих, – і він узгоджується з нашою рекомендацією: за жорсткого бюджету доцільніше вкладатися в адаптацію або екзогенні сигнали умов, ніж у класифікатор сцен за виглядом кадру, який розпізнає критичні умови тільки у 32 % випадків.

Отже, порівняння архітектур детекторів з наявними результатами підтверджує доцільність проведеного саме нашого дослідження: у розглянутих роботах ансамблювання детекторів адаптивне оброблення чи доменна адаптація оцінюються інтегрально, на рівні готової системи, тоді як ізольовального порівняння сценозалежної маршрутизації даних з агностичним ансамблем того самого пулу моделей, кількісного пояснення через кореляцію передбачень та перевірки досяжності знання сцени реальним класифікатором у літературі не наведено – саме ці результати й становлять внесок цієї роботи.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – отримала подальший розвиток методика сценозалежного ансамблевого виявлення об'єктів на аерознімках безпілотників, яка, на відміну від наявних методик, відокремлює внесок сценозалежної маршрутизації даних від внеску ансамблювання детекторів, що дало змогу сформулювати та експериментально підтвердити уточнену умову корисності маршрутизації даних – невідновлюваність сценоспецифічної переваги злиттям повного пулу, показати зумовленість надлишковості передбачень пулу навчальним розподілом даних, а не архітектурою детекторів, та кількісно оцінити систематичні дефекти протоколу оцінювання детекторів на щільних аерознімках.

Практична значущість результатів дослідження – можна використати під час проектування бортових систем виявлення об'єктів на БпЛА з обмеженим обчислювальним ресурсом: за достатнього бюджету доцільно застосовувати глобальний WBF-ансамбль детекторів найсильніших доступних моделей без маршрутизації даних, за жорсткого бюджету за умов доменного зсуву – маршрутизацію за екзогенними сигналами умов (час доби, метеодані з телеметрії), а виправлений протокол оцінювання детекторів застосовний до будь-яких досліджень детекції на щільних аерознімках.

Висновки / Conclusions

Розроблено методику сценозалежного ансамблевого виявлення об'єктів на аерознімках безпілотників, яка відокремлює внесок сценозалежної маршрутизації кадрів від внеску ансамблювання детекторів, що дало змогу сформулювати та експериментально підтвердити уточнену умову корисності маршрутизації даних, показати зумовленість надлишковості передбачень пулу навчальним розподілом даних, а також кількісно оцінити систематичні дефекти протоколу оцінювання детекторів на щільних аерознімках. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Виявлено та виправлено три систематичні дефекти протоколу оцінювання детекторів на щільних аерознімках (поріг впевненості розгортання системи замість оцінювального, невідповідність 10- і 8-класової розмітки, обрізання постоброблення за часом), які спотворювали оцінку окремих моделей до 42 % за mAP@50 і робили некоректним будь-яке порівняння архітектур детекторів; власну реалізацію метрики звірено з еталонною rusocotools (розбіжність $\leq 0,0006$ за mAP@50), а виправлений протокол застосовний до будь-яких досліджень детекції на щільних аерознімках.
2. Встановлено, що ансамбль детекторів WBF трьох детекторів стабільно перевершує найкращу одиночну модель у кожній сцені (+0,051 mAP@50 для базового пулу; ще +0,041 після оновлення пулу чотирма поколіннями YOLO; за класами +0,03...0,08) ціною приблизно триразового зниження швидкості, тобто йдеться про компроміс Парето, а не безумовний виграш.
3. З'ясовано, що в межах досліджених пулів моделей маршрутизація кадру за сценами не перевершує єдиний сценонезалежний ансамбль детекторів (-0,0006 mAP@50, ДІ [-0,0011; -0,0002]) навіть тоді, коли різні сцени віддають перевагу різним сильним моделям; показано, що механізмом є попарне перекриття передбачень 0,78–0,81, однакове для всіх пар поколінь YOLO, тобто у межах перевіреного простору донавчених ваг надлишковість пулу є властивістю навчального розподілу даних, а не архітектурою детектора, а сильних і водночас декорельованих не-YOLO ваг серед публічно доступних станом на липень 2026 р. не знайдено.

4. Сформульовано й підтверджено на двох наборах даних уточнену умову корисності маршрутизації даних: сценоспецифічна перевага має бути невідновлюваною злиттям повного пулу; встановлено, що на наборі даних UAVDT з офіційними мітками умов маршрутизація кадру вперше дає значущий виграш (+0,014 mAP@50 за фіксованого бюджету top-3 ансамблю, крос-фітований відбір), але злиття всіх семи моделей без відбору знову поглинає цю перевагу.
5. Показано, що вузьким місцем розгортання системи є не механізм маршрутизації даних, а знання сцени: класифікатор за виглядом кадру розпізнає туман тільки у 32 % випадків саме в тих умовах, де маршрутизація кадру допомагає, і розгортувана бюджетна ніша зникає; водночас встановлено, що тренування класифікатора, узгоджене з простором міток маршрутизації даних, робить його помилки майже безкоштовними на наборі даних VisDrone.
6. Рекомендовано для практичних систем: за достатнього бюджету – глобальний WBF-ансамбль детекторів найсильніших моделей без маршрутизації даних; за жорсткого бюджету під доменним зсувом – маршрутизацію за екзогенними сигналами умов; перспективою подальших досліджень визначено донавчання сильного крос-парадигмального детектора та перевірку бюджетної ніші без доменного зсуву.

Заява про використання інструментів штучного інтелекту. Під час підготовки цього рукопису автори застосовували інструменти штучного інтелекту та програмне забезпечення для оброблення природної мови (ChatGPT, Grammarly) винятково задля граматичного редагування, коригування орфографії та стилістичної адаптації складних синтаксичних конструкцій. Автори здійснювали критичний аналіз і подальше редагування згенерованого тексту та несуть повну відповідальність за достовірність, науковий зміст і остаточну редакцію публікації.

References

1. Akyon, F. C., Altinuc, S. O., & Temizel, A. (2022). Slicing aided hyper inference and fine-tuning for small object detection. *2022 IEEE International Conference on Image Processing (ICIP)*, 966–970. <https://doi.org/10.1109/ICIP46576.2022.9897990>
2. Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-NMS – Improving object detection with one line of code. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 5561–5569. <https://doi.org/10.1109/ICCV.2017.593>
3. Cao, Y., He, Z., Wang, L., Wang, W., Yuan, Y., Zhang, D., Zhang, J., Zhu, P., Van Gool, L., Han, J., Hoi, S., Hu, Q., & Liu, M. (2021). VisDrone-DET2021: The vision meets drone object detection challenge results. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2847–2854. URL: https://openaccess.thecvf.com/content/ICCV2021W/VisDrone/html/Cao_VisDrone-DET2021_The_Vision_Meets_Drone_Object_Detection_Challenge_Results_ICCVW_2021_paper.html
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *Computer Vision – ECCV 2020*, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13
5. Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. *Computer Vision – ECCV 2018*, 375–391. <https://doi.org/10.48550/arXiv.1804.00518>
6. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., & Wang, Y. (2022). Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7436–7456. <https://doi.org/10.1109/TPAMI.2021.3117837>
7. Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1), 79–87. <https://doi.org/10.1162/neco.1991.3.1.79>

8. Jocher, G., & Qiu, J. (2024). Ultralytics YOLO11 [Computer software]. GitHub. URL: <https://github.com/ultralytics/ultralytics>
9. Jocher, G., Chaurasia, A., & Qiu, J. (2023). *Ultralytics YOLOv8* [Computer software]. GitHub. URL: <https://github.com/ultralytics/ultralytics>
10. Kennerley, M., Wang, J.-G., Veeravalli, B., & Tan, R. T. (2023). 2PCNet: Two-phase consistency training for day-to-night unsupervised domain adaptive object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11484–11493. <https://doi.org/10.48550/arXiv.2303.13853>
11. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. *Computer Vision – ECCV 2014*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
12. Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., & Zhang, L. (2022). Image-adaptive YOLO for object detection in adverse weather conditions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(2), pp. 1792–1800. <https://doi.org/10.1609/aaai.v36i2.20072>
13. Ouyang-Zhang, J., Cho, J. H., Zhou, X., & Krähenbühl, P. (2022). NMS strikes back. *arXiv*. <https://doi.org/10.48550/arXiv.2212.06137>
14. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788. <https://doi.org/10.1109/CVPR.2016.91>
15. Sakseena, S. K. (2025). *VisDrone detection model zoo* [Model collection]. Hugging Face. URL: <https://huggingface.co/collections/dronefreak/visdrone-detection-model-zoo>
16. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *International Conference on Learning Representations (ICLR)*, pp. 1–19. <https://doi.org/10.48550/arXiv.1701.06538>
17. Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, article ID 104117. <https://doi.org/10.1016/j.imavis.2021.104117>
18. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37 (NeurIPS 2024), article ID 93301. <https://doi.org/10.48550/arXiv.2405.14458>
19. Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. *Computer Vision – ECCV 2024*, 1–21. <https://doi.org/10.48550/arXiv.2402.13616>
20. Zhang, H., Liu, K., Gan, Z., & Zhu, G.-N. (2025). UAV-DETR: Efficient end-to-end object detection for unmanned aerial vehicle imagery. *arXiv*. <https://doi.org/10.48550/arXiv.2501.01855>
21. Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16965–16974. <https://doi.org/10.48550/arXiv.2304.08069>
22. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2021). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7380–7399. <https://doi.org/10.1109/TPAMI.2021.3119563>
23. Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2778–2788. <https://doi.org/10.48550/arXiv.2108.11539>

Kh. S. Ruda, V. O. Kromkach

Lviv Polytechnic National University, Lviv, Ukraine

SCENE-DEPENDENT ENSEMBLE OBJECT DETECTION ON UAV AERIAL IMAGERY: THE LIMITS OF USEFULNESS OF SCENE-BASED FRAME ROUTING

Scene-dependent ensemble object detection on unmanned aerial vehicle imagery is investigated, in which a scene classifier (day/night/rain) routes each frame to a specialized per-scene ensemble of detectors whose predictions are merged by weighted boxes fusion (WBF). An isolating analysis is performed that separates the contribution of scene-based routing from the contribution of ensembling itself; no new architecture is proposed. Seven fine-tuned detectors spanning five YOLO generations are evaluated per scene under a unified protocol, with paired and cluster bootstrap significance testing and zero-shot transfer to a second dataset. Three systematic defects of the evaluation protocol are identified and corrected – scoring at the deployment confidence threshold, a 10-versus-8-class label-layout mismatch, and a time-limited post-processing path – which distort per-model scores by tens of percent. It is established that, under the corrected protocol, a WBF ensemble of the three strongest detectors consistently outperforms the best single model in every scene (by about +0.05 mAP@50) at roughly a threefold cost in inference speed, which constitutes a Pareto trade-off. It is found that scene-based frame routing does not outperform a single scene-agnostic ensemble of the same models (a difference of –0.0006 mAP@50), and this result is reproduced on a heterogeneous pool: the redundancy of pool predictions is shown to be a property of the training data distribution rather than of the architecture, within the space of public fine-tuned weights available as of July 2026. It is demonstrated that on UAVDT, with official condition labels (day/night/fog), routing yields a statistically significant gain for the first time under a fixed ensemble compute budget (+0.014 mAP@50), yet fusing the entire pool without any selection absorbs this advantage again. A refined condition for the usefulness of routing is formulated and confirmed on two datasets: the scene-specific advantage must be unrecoverable by fusing the full pool; the practical niche of routing is shown to be a hard compute budget under domain shift with an exogenous source of scene knowledge, since an appearance-based classifier recognizes the critical adverse conditions (fog) in only 32 % of cases.

Keywords: weighted boxes fusion; deep learning; convolutional neural networks; VisDrone; UAVDT; domain shift.