



DOI 10.28925/2663-4023.2026.34.1340

УДК 004.032.26:004.932

Руда Христина Степанівна

доктор філософії, доцент кафедри захисту інформації

Національний університет «Львівська політехніка», Львів, Україна

ORCID:0000-0001-8644-411X

Khrystyna.S.Ruda@lpnu.ua

Кромкач Владислав Олександрович

асистент кафедри захисту інформації

Національний університет «Львівська політехніка», Львів, Україна

ORCID:0009-0001-5608-5715

vlad.kromkach@gmail.com

АДАПТИВНИЙ КАСКАД ЗА СКЛАДНІСТЮ СЦЕНИ ДЛЯ ВИЯВЛЕННЯ ОБ'ЄКТІВ НА АЕРОЗНІМКАХ БЕЗПІЛОТНИКІВ: КОЛИ ВИБІРКОВА ЕСКАЛАЦІЯ ОБЧИСЛЕНЬ ВИПРАВДАНА

Анотація. Безпілотні літальні апарати є штатним інструментом відеомоніторингу в задачах безпеки – від охорони периметра й критичної інфраструктури до ситуаційної обізнаності, – а ансамблі детекторів підвищують точність виявлення об'єктів на аерознімках ціною кратно більших обчислень на кожному кадрі, що критично для бортових систем. Запропоновано й досліджено адаптивний каскад за складністю сцени (Scene-Complexity Adaptive Cascade, SCAC): дешевий детектор першого ярусу обробляє кожен кадр і обчислює безрозмірний сигнал складності без використання розмітки; лише кадри зі складністю понад поріг ескалюються на дорогий WBF-ансамбль другого ярусу. Поріг не підбирається вручну – він калібрується крос-фітуванням на рівні відеопослідовностей під цільовий середній бюджет. Методологічним внеском є оракульний гейт придатності: перш ніж будувати метод, за кешованими передбаченнями будується жадібна референсна крива за реалізованим перекладним приростом ансамблю, з якою порівнюються реалізовані сигнали складності. Гейт дає негативну частину результату: у навчальному домені (VisDrone-DET) приріст ансамблю щільний за кадрами (покращується 75-90 % кадрів), тож реалізовані сигнали складності не перевершують рівномірне злиття за того самого бюджету (селективний потенціал, який виявляє лише оракул з доступом до розмітки, для них недосяжний), а поширений сигнал невпевненості стабільно гірший за випадкову ескалацію. Натомість за доменного зсуву (нульове перенесення на UAVDT з офіційними мітками умов) приріст ансамблю розріджений (41 % кадрів; уночі ансамбль шкодить), і крос-фітований каскад із сигналом щільності детекцій досягає якості повного ансамблю (0,618 mAP@50, збалансовано за сценами) за 45 % його обчислень, значуще перевершуючи одиничну модель (+0,022 на повній вибірці; бутстреп: середнє +0,017, 95 % ДІ [+0,002; +0,032]) та випадкову ескалацію, зокрема стратифіковану за сценами (+0,019 на повній вибірці; бутстреп: середнє +0,016, 95 % ДІ [+0,001; +0,029]). Реалізований розподіл ескалації (9 % ніч / 22 % день / 45 % туман) відтворює сценозалежність без класифікатора сцен, обминаючи відоме вузьке місце маршрутизації – нерозпізнаваність несприятливих умов за виглядом кадру. Сформульовано практичний критерій: вибіркова ескалація обчислень виправдана лише там, де приріст дорогої моделі розріджений за кадрами, – типово за доменного зсуву, а не в навчальному домені.

Ключові слова: виявлення об'єктів; аерознімки БпЛА; каскад моделей; ансамбль детекторів; адаптивні обчислення; доменний зсув; VisDrone; UAVDT.

ВСТУП

Безпілотні літальні апарати (БпЛА) стали штатним інструментом відеомоніторингу в задачах безпеки: охорона периметра та об'єктів критичної інфраструктури, ситуаційна обізнаність, документування наслідків надзвичайних ситуацій. Автоматичне виявлення об'єктів на аерознімках у таких застосуваннях поєднує три труднощі: дрібні об'єкти, високу щільність і мінливі умови зйомки (день, ніч, туман). Ансамблювання кількох детекторів зі злиттям рамок методом WBF [1] стабільно



підвищує точність: за результатами цієї роботи на VisDrone ансамбль із трьох моделей досягає $\approx 0,535$ mAP@50 проти 0,480 у найкращій одиничній моделі, тобто +0,055 (рис. 1, середня панель), – але з майже пропорційною ціною: 2290 проти 684 GFLOPs/кадр, тобто приблизно у 3,35 рази більше обчислень (див. «Експериментальна постановка»). Для бортових і потокових систем моніторингу, де обчислювальний ресурс та енергоживлення обмежені, це часто неприйнятно.

Постановка проблеми. Природна ідея – витрачати обчислення нерівномірно: «легкі» кадри закривати дешевою моделлю, «складні» – ескалювати на ансамбль. Такі каскади добре відомі, однак у літературі про детекцію на аерознімках бракує перевірки передумови каскаду: чи існують взагалі кадри, на яких дорога модель не потрібна? У [2] показано, що на осі ідентичності моделі (сценозалежна маршрутизація «яку модель застосувати») адаптивність у навчальному домені не окупається: злиття повного пулу поглинає сценоспецифічні переваги. Ця робота досліджує ортогональну вісь – скільки обчислень витрачати на кадр – і показує, що вона підкоряється тій самій емпіричній закономірності, з тією самою межею і тим самим винятком.

Аналіз останніх досліджень і публікацій. Каскадна фільтрація сягає детектора Віоли–Джонса [3]; у нейронних мережах їй відповідають архітектури з раннім виходом (BranchyNet [4]) та дворівневі схеми «дешева/дорога модель» у відеоаналітиці (NoScore [5], Chameleon [6], FrameExit [7]). Сучасні огляди систематизують ранній вихід (early exit) і розділені обчислення для крайових пристроїв [8], ширший клас динамічних нейронних мереж, які адаптують обчислення до кожного входу [9], та архітектури з раннім виходом як окремий напрям [10]; у них зазначено, що виграш умовних обчислень виникає лише тоді, коли для частини входів обчислення надлишкові, – саме цю передумову пропонований гейт перевіряє явно. Відмінність постановки цієї роботи потрібна: по-перше, другий ярус – не одна велика модель, а WBF-ансамбль; по-друге, порівняння ведеться за вирівняним середнім бюджетом проти найсильніших рівномірних конфігурацій (одиничних моделей, пар, повного ансамблю), а не лише проти дорогої моделі; по-третє, придатність каскаду перевіряється оракульним гейтом до побудови методу.

Метод зваженого злиття рамок WBF [1] усереднює збіжні рамки з вагами за впевненістю і є стандартом для ансамблів детекторів; передумовою корисності є декореляція помилок членів. Виміряне у [2] попарне перекриття передбачень публічних детекторів, донавчених на одному наборі даних, становить 0,78-0,81 незалежно від архітектури – надлишковість, що обмежує і ансамбль, і маршрутизацію. Умовна маршрутизація обчислень походить від сумішей локальних експертів [11] і розрізаних архітектур Mixture-of-Experts [12]; на відміну від маршрутизації за типом входу (сцени), пропонований каскад маршрутизує за складністю, вимірюваною з виходу самого детектора першого ярусу, – сигнал доступний безкоштовно і не потребує окремого класифікатора. У детекції на аерознімках останніми роками домінують спеціалізовані архітектури для дрібних об'єктів – від TPH-YOLOv5 з трансформерними головами передбачень [13] до адаптації UAV-DETR [14]. Українські дослідники активно застосовують моделі YOLO до розпізнавання об'єктів у реальному часі: у [15] побудовано систему розпізнавання на основі YOLOv8 для зашумленого та змінного середовища, а в [16] запропоновано адаптивну модифікацію YOLOv11 з механізмами уваги для розпізнавання техніки у відеопотоці. Проте ані питання обчислювального бюджету на кадр, ані передумова каскаду (існування кадрів, на яких дорога конфігурація не потрібна) у цих роботах не розглядаються. Еталонними наборами даних слугують VisDrone-DET [17] (навчальний домен) і UAVDT [18] (зовнішня валідація з офіційними мітками умов день / ніч / туман).

Отже, не вирішеними частинами загальної проблеми залишаються: відсутність дешевого способу оцінити доцільність каскаду до його побудови; невідомість умов, за яких вибіркова ескалація обчислень для детекції на аерознімках взагалі окупається; та відсутність порівнянь каскадів із найсильнішими рівномірними конфігураціями за вирівняним бюджетом. Розмежування з [2]: там досліджено вісь ідентичності моделі – яку конфігурацію застосувати залежно від сцени; ця робота досліджує ортогональну вісь – обсяг обчислень на кадр; спільними є пул моделей і протокол оцінювання, тоді як метод, експерименти та всі результати – нові.

Мета статті. Розробити та експериментально дослідити адаптивний каскад за складністю сцени для виявлення об'єктів на аерознімках БпЛА, який калібрує поріг ескалації без ручного налаштування, і сформулювати умови, за яких вибіркова ескалація обчислень виправдана.

Наукова новизна одержаних результатів. (1) Запропоновано оракульний гейт придатності – жадібну референсну криву за реалізованим пер-кадровим природним ансамблем та її блокове уточнення за пул-рівневою метрикою, які можна побудувати з кешованих передбачень до реалізації методу і використати як дешевий перший крок проєктування адаптивних детекційних систем. (2) Одержано двосторонній результат: у навчальному домені ані референсна крива, ані реалізовані сигнали не перевершують рівномірне злиття за рівного бюджету, а за доменного зсуву крос-фітований каскад значуще перевершує одиничну модель і обидва випадкові контролю, досягаючи якості повного ансамблю



за 45 % його обчислень. (3) Показано емерджентну сценорадаптивність без знання сцени: сигнал щільності детекцій самостійно концентрує ескалацію в тумані та майже вимикає її вночі. (4) Встановлено, що сигнал невпевненості – найінтуїтивніший вибір – стабільно не кращий за випадкову ескалацію на обох наборах даних.

МЕТОДИКА ДОСЛІДЖЕННЯ

Двоярусний каскад. Ярус 1 – одинична модель – обробляє кожен кадр потоку. З її виходу обчислюється сигнал складності $c(x)$ (без розмітки, без додаткової моделі). Рішення каскаду ухвалюється пороговим правилом (1):

$$y(x) = \text{WBF}(\text{ярус 1} \cup \text{ярус 2})(x), \text{ якщо } c(x) \geq \tau; \text{ інакше } y(x) = \text{ярус 1}(x) \quad (1)$$

тобто на ескальованому кадрі виконуються решта членів ансамблю, і всі передбачення (включно з ярусом 1) зливаються WBF; інакше відповіддю стають детекції ярусу 1. Середній обчислювальний бюджет на кадр визначається виразом (2):

$$B(p) = B_1 + p \cdot B_2, \quad (2)$$

де p – частка ескальованих кадрів, B_1 – вартість ярусу 1, B_2 – вартість додаткових членів (передбачення ярусу 1 повторно використовуються у злитті, тому, коли ярус 1 сам є членом ансамблю, ескалація доплачує лише за інших). Сигнали складності для порівняння: C-count – кількість детекцій ярусу 1 із впевненістю не нижче 0,25; C-conf – одиниця мінус середня впевненість топ-10 детекцій; C-entropy – середня бінарна ентропія впевненостей. Вибір порога τ не ручний: на калібрувальних даних обирається частка p^* , що максимізує збалансований за сценами mAP@50, і τ встановлюється як відповідний квантиль сигналу.

Оракульний гейт придатності. Перш ніж реалізовувати каскад, за кешованими передбаченнями обох ярусів обчислюється реалізований пер-кадровий приріст AP кожного кадру, і будується жадібна референсна крива: кадри ескалюються в порядку спадання цього приросту. Слід підкреслити, що ця крива не є строгою верхньою межею цільової метрики: mAP обчислюється на спільному пулі детекцій усіх кадрів, тому внесок кадру не адитивний – ескалація одного кадру зсуває позиції детекцій інших кадрів у глобальному ранжуванні за впевненістю. Жадібне впорядкування за пер-кадровим приростом оптимізує сепарабельний сурогат метрики, а збалансоване за сценами усереднення додає ще одну нелінійність (вага кадру залежить від наповненості його сцени). Тому додатково побудовано блокову жадібну криву за самою пул-рівневою метрикою: кадри ескалюються пачками по 20, і кожна пачка обирається за фактичним маргінальним приростом цільового mAP@50 з повним перерахунком метрики після кожного блоку. Блокова крива суттєво вища за пер-кадрову і ближча до справжньої межі, хоча також нею не є (жадібний вибір не гарантує оптимуму) і, як і пер-кадрова, використовує розмітку, тобто є оракульним референсом, а не методом. Побудова обох кривих потребує лише злиття кешованих передбачень (жодного інференсу).

Коректний вибір порога та статистика. На UAVDT поріг калібрується двофолдовим крос-фітуванням (перехресним допасуванням, cross-fitting) на рівні відеопослідовностей – покадровий поділ дав би витік даних через майже-дублікати сусідніх кадрів: послідовності стратифіковано за сценами діляться на дві частини (фолди) A/B; поріг обирається на одній частині й застосовується до іншої, тож рішення про ескалацію кожного кадру ухвалює поріг, на який його послідовність не впливала. Довірчі інтервали – парний кластерний бутстреп (bootstrap; 1000 реплік, ресемплінг послідовностей у межах сцени; всі конфігурації оцінюються на спільному ресемплі). Контролі: (а) глобальна випадкова ескалація тієї самої загальної частки – контроль, зареєстрований протоколом заздалегідь; (б) стратифікована випадкова ескалація з точно тими самими пер-сеновими кількостями ескальованих кадрів – суворіший контроль, що ізолює внутрішньосенову інформативність сигналу від міжсенового розподілу бюджету. Випадкові контролі використовують окремий генератор випадковості, щоб їх додавання не змінювало жодних інших чисел. Наведені довірчі інтервали є маргінальними: вони не враховують ані множинності п'яти порівнянь, ані відбору сигналу C-count з-поміж трьох досліджених на тих самих даних.

Експериментальна постановка. Використано протокол оцінювання, описаний у [2]: поріг впевненості оцінювання 0,001 (не робочий поріг розгортання 0,25), роздільна здатність 1280 для точності, WBF з IoU 0,55, власна реалізація mAP звірена з еталонною русocotools (розбіжність не перевищує 0,0006 за mAP@50 та 0,0016 за mAP@50-95). Пул моделей – публічні донавчені на VisDrone ваги п'яти поколінь YOLO з відкритого модельного зоопарку [19]: YOLOv8 (m/l/x), YOLOv9e [20], YOLOv10x [21], YOLO11x [22], YOLO26x; чотири новіші покоління треновані за єдиним рецептом (300

епох, роздільна здатність 640). Ярус 2 – найкращий сценонезалежний ансамбль {YOLOv9e, YOLO26x, YOLOv10x}. Бюджети (GFLOPs/кадр на роздільній здатності 1280): ярус 1 YOLOv10x – 684; ескалація (YOLOv9e + YOLO26x) – 1606; трійка – 2290; найкраща пара – 1455. VisDrone: 441 денне / 542 нічні / 91 дощове оцінювальне зображення; офіційних міток умов VisDrone не має, тож розбиття за сценами виконано евристичним класифікатором за ознаками яскравості, насиченості та контрасту, описаним у [2]; нічні зображення взято з тестового розбиття, оскільки валідаційне містить лише 16 нічних зображень, – усі порівнювані конфігурації оцінюються на тих самих кадрах, тож зсуву порівнянь це не створює. UAVDT: нульове перенесення, 1372 кадри з 50 послідовностей (29 день / 13 ніч / 8 туман, офіційні мітки); 8-класовий простір передбачень пулу зведено до 3-класового простору UAVDT – фургони (van) зліто з легковими автомобілями (car) відповідно до визначення категорії car в UAVDT, вантажівки й автобуси відображено безпосередньо, решту класів (пішоходи, велосипеди, мотоцикли тощо) відкинута; рамки в офіційних ignore-зонах відкинута за центром. GFLOPs обчислено профілювальником на роздільній здатності 1280; латентність виміряно на Apple M-series (MPS). Код експериментів і кешовані передбачення доступні від авторів за запитом.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ ТА ЇХ ОБГОВОРЕННЯ

У навчальному домені гейт закриває каскад (VisDrone). Приріст ансамблю над ярусом 1 щільний за кадрами: ярус 2 покращує 90 % кадрів при дешевому ярусі 1 (YOLOv8s) і 75 % – при сильному (YOLOv10x). «Легких» кадрів, на яких можна заощадити, майже немає, тому (рис. 1): з ярусом 1 = YOLOv8s навіть жадібна референсна крива програє сучасним одиничним моделям за рівного бюджету (при бюджеті YOLO26x, 835 GFLOPs: $\approx 0,464$ проти $0,4795$ у самої YOLO26x); з ярусом 1 = YOLOv10x референсна крива лише трохи перевищує рівномірне злиття пар ($+0,004 \dots 0,010$ залежно від пари за вирівняного бюджету – величини порядку статистичного шуму), а найкращий реалізований сигнал C-count іде з парами практично нарівні (від $-0,002$ до $+0,004$); сигнал C-conf гірший за випадкову ескалацію на всій кривій – кадри з низькою впевненістю виявляються розрідженими кадрами з малою вагою в метриці. Блокова жадібна крива за пул-рівневою метрикою уточнює цей вердикт: селективний потенціал у домені існує в принципі – за 1073 GFLOPs/кадр (частка ескалації $\approx 0,24$) вона досягає якості пар ($0,516$ проти $0,511-0,517$), а за 1671 GFLOPs (73 % бюджету трійки) – $0,537$, дещо вище повної трійки ($0,535$). Проте цей потенціал видно лише оракулу, який використовує розмітку: реалізовані сигнали йдуть уздовж пер-кадрової кривої ($0,501-0,508$ при $p = 0,2-0,3$), глибоко нижче блокової. Відтак негативний висновок для навчального домену стосується саме реалізованих сигналів складності, а не принципової відсутності селективного потенціалу. Це обчислювальний аналог результату [2] про маршрутизацію за сценами: коли приріст дорогої конфігурації щільний за кадрами, рівномірне застосування поглинає будь-яку реалізовану вибіркковість.

На рис. 1 ліва й середня панелі відповідають VisDrone (зважене за кількістю зображень mAP@50; ярус 1 – YOLOv8s та YOLOv10x відповідно), права – UAVDT (збалансоване за сценами mAP@50, нульове перенесення); зіркою позначено крос-фітований каскад, штрих-пунктирною кривою – блоковий жадібний оракул за пул-рівневою метрикою; сигнал C-ентропу на рис. 1 опущено для розбірливості – він лежить між C-count і C-conf (див. рис. 2).

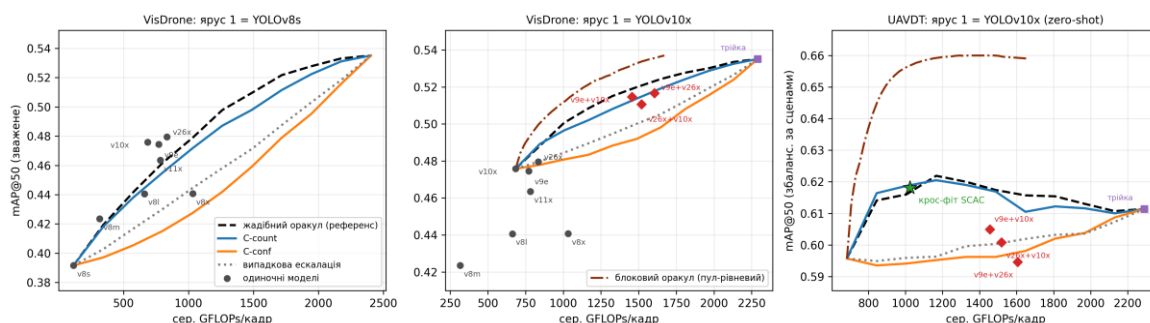


Рис. 1. Оракульний гейт: mAP@50 проти середнього бюджету, GFLOPs/кадр (ліва й середня панелі – VisDrone, права – UAVDT)

За доменного зсуву каскад виграє (UAVDT). На UAVDT картина протилежна: трійка покращує лише 41 % кадрів, а вночі шкодить ($0,6086$ проти $0,6194$ у одниничної YOLOv10x). Крива «mAP проти частки ескалації» стає немонотонною з піком при $p \approx 0,2-0,3$ (рис. 2) – ескалація понад пік повертає в ансамбль кадри, де він гірший. На ділянці $p \approx 0,1-0,2$ крива C-count проходить навіть вище пер-кадрової

референсної кривої (рис. 1 і 2, праві панелі) – це можливо саме тому, що пер-кадрове впорядкування не є верхньою межею пул-рівневої метрики; блокова жадібна крива підтверджує це кількісно: її пік 0,660 при $p \approx 0,38$ значно вищий і за пер-кадрову криву (пік 0,622), і за реалізований каскад (0,618), тобто пер-кадрове впорядкування суттєво недооцінює справжню межу. Крос-фітований каскад (обрані фолдами частки 0,25 і 0,15; реалізована частка 21,1 %) досягає якості повного ансамблю за 45 % його обчислень (табл. 1); статистичні гарантії наведено в табл. 2: каскад значуще перевершує одиничну модель і обидва випадкові контролю (отже, сигнал інформативний і всередині сцен, а не лише через міжсценний розподіл бюджету), а від пари та трійки статистично не відрізняється – за 70 % та 45 % їхнього бюджету відповідно.

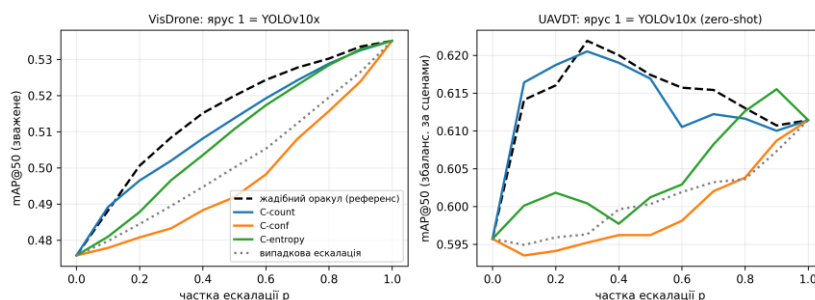


Рис. 2. Чутливість до порога: $mAP@50$ як функція частки ескалації p (ліворуч – VisDrone, зважене $mAP@50$; праворуч – UAVDT, збалансоване за сценами)

Таблиця 1

Конфігурації на UAVDT: якість проти середнього бюджету

Конфігурація	GFLOPs/кадр	$mAP@50$ (збаланс.)	$mAP@50$ (зваж.)
Одинична YOLOv10x	684	0,5957	0,6058
Каскад SCAC (крос-фіт)	1024	0,6181	0,6330
Найкраща пара {9e, 10x}	1455	0,6049	0,6285
Повна трійка	2290	0,6114	0,6271

Таблиця 2

Різниця збалансованого за сценами $mAP@50$ «каскад – конфігурація», UAVDT

Різниця (каскад –)	Δ (повна вибірка)	Середнє бутстрепу [95 % ДІ]	Частка реплік $\Delta \leq 0$
одинична YOLOv10x	+0,0224	+0,0171 [+0,0021; +0,0318]	0,012
випадкова ескалація (глобальна)	+0,0212	+0,0169 [+0,0023; +0,0308]	0,010
випадкова ескалація (стратифікована)	+0,0187	+0,0158 [+0,0014; +0,0287]	0,014
найкраща пара {9e, 10x}	+0,0132	+0,0080 [–0,0102; +0,0219]	0,153
повна трійка	+0,0067	+0,0037 [–0,0167; +0,0197]	0,294

Примітка. Точкова оцінка – різниця збалансованих $mAP@50$ на повній вибірці (узгоджена з табл. 1; для випадкових контролів – середнє трьох розіграшів). Інтервал – перцентильний кластерний бутстреп; середнє бутстреп-розподілу систематично менше за повновибіркову різницю через нелінійність збалансованої статистики за ресемплінгу послідовностей (зсув бутстреп-середнього). Остання колонка – частка бутстреп-реплік із недодатною різницею, а не p -значення у частотному сенсі.

Реальна латентність і апаратні аспекти. Латентність вимірюно одним і тим самим вимірювачем на Apple M-series (MPS): на роздільній здатності розгортання 640 одинична модель коштує 66 мс/кадр, трійка – 249 мс/кадр; середня латентність каскаду, складена з цих вимірів за реалізованої частки ескалації 21,1 %, становить 105 мс/кадр (9,5 кадр/с; у 2,4 раза швидше за трійку). На оцінювальній роздільній здатності 1280 – 198, 912 і 349 мс відповідно (у 2,6 раза швидше за трійку). Розподіл часу обробки при цьому бімодальний: медіана дорівнює латентності ярусу 1 (66 мс на 640; 198 мс на 1280), тоді як 95-й і 99-й перцентилі – латентності повної трійки (249 мс; 912 мс), оскільки ескалюється 21,1 % кадрів; для потокової системи з жорстким дедлайном на кадр критичним є саме гірший випадок, а не



середнє. Каскад зменшує середню кількість операцій, але не резидентний обсяг пам'яті: рішення про ескалацію ухвалюється під час обробки кадру, тож усі три моделі мають бути завантажені одночасно – 112 МБ (YOLOv9e) + 61 МБ (YOLOv10x) + 113 МБ (YOLO26x), сумарно ≈ 286 МБ ваг без урахування активацій; для вбудованих платформ, де пам'ять часто є визначальним обмеженням, виграш каскаду лежить в енергії та пропускній здатності, а не в пам'яті. Зауваження щодо повноти моделі вартості: вимір трійки включає накладні витрати WBF-злиття, однак GFLOPs-бюджет не враховує вартості самого злиття, перемикання моделей і передавання даних; вимірювальна платформа (MPS) не є цільовою бортовою, тож абсолютні часи слід розглядати як відносні орієнтири.

Емерджентна сценорадаптивність. Реалізована частка ескалації становить 8,8 % уночі, 21,9 % удень і 45,1 % у тумані. Сигнал щільності детекцій – без жодного знання про сцену – концентрує обчислення там, де ансамбль допомагає (туман), і майже вимикає їх там, де він шкодить (ніч). Це розв'язує проблему, показану в [2] для маршрутизації за сценами: класифікатор умов за виглядом кадру розпізнає туман лише у 32 % випадків, і бюджетна ніша маршрутизації зникала саме через це. Каскадові класифікатор не потрібен: складність вимірюється з виходу самого детектора.

Підсумки абляційних експериментів. Сигнал: C-count наближається до референсної кривої на малих частках ескалації; C-entropy посередині; C-conf не кращий за випадкову ескалацію на всій кривій VisDrone і на малих та середніх частках на UAVDT – головний контрінтуїтивний висновок для практиків. Чутливість до порога: крива гладка; на UAVDT – з піком, який крос-фітування знаходить без витоку даних. Склад ярусу 2: заміна ансамблю однією великою моделлю неможлива без втрат – жодна одинична модель не досягає якості трійки в тумані; водночас пара {YOLOv9e, YOLOv10x} закриває більшу частину розриву за 64 % бюджету трійки, тож «ярус 2 = пара» – практична опція для найжорсткіших бюджетів. Триярусний каскад: попередня оцінка структури «одинична модель → пара → трійка» з двома крос-фітованими порогоми на тому самому сигналі не дає виграшу над двоярусним каскадом ($-0,0003$ збалансованого mAP@50 за повною вибіркою; частка реплік $\Delta \leq 0 - 0,45$) за помітно більшого середнього бюджету (1619 проти 1024 GFLOPs/кадр): проміжний ярус збільшує вартість, не додаючи якості, – двоярусної структури достатньо.

Стійкість. Два додаткові експерименти (табл. 3) перевіряють чутливість методу до довільних виборів. По-перше, поріг видимості детекцій у сигналі C-count: за порогів 0,1 та 0,5 замість базового 0,25 якість каскаду практично не змінюється (0,618 за всіх трьох порогів), повновибіркова перевага над одиничною моделлю становить $+0,0224 \dots +0,0227$, над стратифікованим контролем $+0,0187 \dots +0,0194$, і всі бутстреп-інтервали цих різниць виключають нуль (частка реплік $\Delta \leq 0$ не перевищує 0,021) – сигнал щільності нечутливий до цього гіперпараметра. По-друге, вибір ярусу 1: із заміною YOLOv10x на YOLO26x – модель, слабшу в цільовому домені, – приріст трійки над ярусом 1 стає щільним, і крос-фітування коректно обирає майже повну ескалацію (84 % кадрів, 90 % бюджету трійки): каскад вироджується в рівномірний ансамбль, бюджетний виграш зникає, і перевага над випадковими контролями втрачає значущість (перевага над одиничною YOLO26x при цьому лишається значущою, $+0,0355$ за повною вибіркою). Це не збій, а дія сформульованої закономірності всередині самого методу: коли приріст щільний, заощаджувати нема на чому, і калібрування виявляє це самостійно. Практичний висновок: ярусом 1 має бути найсильніша доступна в цільовому домені дешева модель.

Таблиця 3

Стійкість каскаду до порога сигналу та вибору ярусу 1 (UAVDT; Δ – збалансований mAP@50)

Варіант	Частка ескалації	GFLOPs/кадр	mAP@50 (збаланс.)	Δ до одиничної: повна вибірка (бутстреп [95 % ДІ])	Δ до стратиф. контролю: повна вибірка (бутстреп [95 % ДІ])
Сигнал при порозі 0,1	0,191	991	0,6184	$+0,0227 (+0,0169 [+0,0009; +0,0332])$	$+0,0194 (+0,0149 [+0,0005; +0,0286])$
Основний (поріг 0,25)	0,211	1024	0,6181	$+0,0224 (+0,0171 [+0,0021; +0,0318])$	$+0,0187 (+0,0158 [+0,0014; +0,0287])$
Сигнал при порозі 0,5	0,366	1272	0,6181	$+0,0224 (+0,0173 [+0,0013; +0,0339])$	$+0,0189 (+0,0153 [+0,0007; +0,0286])$
Ярус 1 = YOLO26x	0,842	2060	0,6091	$+0,0355 (+0,0309 [+0,0053; +0,0640])$	$+0,0047 (+0,0035 [-0,0091; +0,0116])$

Примітка. Схема подання, як у табл. 2: перше число – різниця на повній вибірці, у дужках – середнє бутстреп-розподілу з перцентильним 95 % довірчим інтервалом.



Обмеження. Оцінка на UAVDT – нульове перенесення (моделі, донавчені в цільовому домені, могли б змінити баланс); туман представлений лише 8 послідовностями; латентність каскаду складена з вимірів компонентів на одному кадрі після програвання, а не виміряна окремим прогоном, чутлива до теплового стану пристрою і отримана на платформі (MPS), що не є цільовою бортовою. Жадібні референсні криві (пер-кадрова та блокова пул-рівнева) не є гарантованими верхніми межами метрики: перша оптимізує сепарабельний сурогат, друга не гарантує оптимуму через жадібний вибір, і обидві використовують розмітку; негативний вердикт гейта на VisDrone стосується реалізованих сигналів, які йдуть уздовж пер-кадрової кривої й не наближаються до блокової. Блокова крива, до того ж, обирає блоки за фактичним приростом цільової метрики на тих самих даних оцінювання, тож частина розриву між нею (0,660 на UAVDT) і реалізованим каскадом (0,618) є переобладнанням оракула під вибірку, а не досяжним потенціалом. Виявлену закономірність слід розуміти як емпіричну: її підтверджено на двох наборах даних, одному пулі моделей однієї архітектурної родини (YOLO) та одній задачі.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Встановлено, що вибіркова ескалація обчислень підкоряється тій самій емпіричній закономірності, що й маршрутизація за сценами: адаптивність окупається лише там, де перевага дорогої конфігурації розріджена за кадрами. У навчальному домені приріст ансамблю щільний – і реалізовані сигнали складності не перевершують рівномірне злиття за рівного бюджету, хоча блоковий пул-рівневий оракул показує, що принциповий селективний потенціал існує (0,537 проти 0,535 у трійки за 73 % її бюджету), – він просто недосяжний без розмітки. За доменного зсуву приріст розріджений і частково від’ємний – і простий каскад із сигналом кількості детекцій, коректно каліброваним порогом і жодним знанням про сцену досягає якості повного ансамблю за 45 % його обчислень, значуще перевершуючи одиничну модель і випадковий контроль; середня латентність при цьому у 2,4 раза нижча за латентність трійки (складена оцінка з вимірів компонентів, а не окремий прогін каскаду). Запропонований оракульний гейт придатності рекомендовано як перший крок побудови адаптивних детекційних систем: він потребує лише злиття кешованих передбачень і або збереже місяці розробки, або дасть їй обґрунтування. Перспективи подальших досліджень: перенесення та вимірювання на вбудованій платформі класу NVIDIA Jetson із профілюванням пам’яті й енергоспоживання; перевірка каскаду з пулом, донавченим у цільовому домені (чи зберігається ніша без доменного зсуву); реалізовані сигнали, що наближаються до блокової пул-рівневої кривої; поєднання каскаду з екзогенними сигналами умов у бортових системах моніторингу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 104117. <https://doi.org/10.1016/j.imavis.2021.104117>
2. Ruda, Kh. S., & Kromkach, V. O. (in press). Scene-dependent ensemble object detection in UAV aerial imagery: Limits of the utility of scene-based routing. *Scientific Bulletin of UNFU*.
3. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 511-518). <https://doi.org/10.1109/CVPR.2001.990517>
4. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). BranchyNet: Fast inference via early exiting from deep neural networks. In *23rd International Conference on Pattern Recognition (ICPR)* (pp. 2464-2469). <https://doi.org/10.1109/ICPR.2016.7900006>
5. Kang, D., Emmons, J., Abuzaid, F., Bailis, P., & Zaharia, M. (2017). NoScope: Optimizing neural network queries over video at scale. *Proceedings of the VLDB Endowment*, 10(11), 1586-1597. <https://doi.org/10.14778/3137628.3137664>
6. Jiang, J., Ananthanarayanan, G., Bodik, P., Sen, S., & Stoica, I. (2018). Chameleon: Scalable adaptation of video analytics. In *Proceedings of the ACM SIGCOMM 2018 Conference* (pp. 253-266). <https://doi.org/10.1145/3230543.3230574>
7. Ghodrati, A., Bejnordi, B. E., & Habibian, A. (2021). FrameExit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15608-15618). <https://doi.org/10.1109/CVPR46437.2021.01535>
8. Matsubara, Y., Levorato, M., & Restuccia, F. (2022). Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys*, 55(5), 1-30. <https://doi.org/10.1145/3527155>



9. 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>
10. Rahmath P., H., Srivastava, V., Chaurasia, K., Pacheco, R. G., & Couto, R. S. (2024). Early-exit deep neural network – A comprehensive survey. *ACM Computing Surveys*, 57(3), 1-37. <https://doi.org/10.1145/3698767>
11. 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>
12. 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. In *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1701.06538>
13. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2778-2788). <https://doi.org/10.1109/ICCVW54120.2021.00312>
14. 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>
15. Nazarkevych, M. A., & Oleksiv, N. T. (2024). Object recognition system based on the YOLO model. *Ukrainian Journal of Information Technology*, 6(1), 120-126. <https://doi.org/10.23939/ujit2024.01.120>
16. Filimonchuk, T. V., Koltun, Yu. M., & Maslov, M. K. (2026). Adaptive model for equipment recognition with attention mechanisms based on YOLOv11. *Information and Control Systems for Railway Transport*, 31(2). <https://doi.org/10.18664/iksz.v31i2.362205>
17. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022). 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>
18. 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. In *Computer Vision – ECCV 2018* (pp. 370-386). https://doi.org/10.1007/978-3-030-01249-6_23
19. Saksena, S. K. (2025). *VisDrone detection model zoo* [Model collection]. Hugging Face. <https://huggingface.co/collections/dronefreak/visdrone-detection-model-zoo>
20. Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision – ECCV 2024* (Lecture Notes in Computer Science, Vol. 15089, pp. 1–21). https://doi.org/10.1007/978-3-031-72751-1_1
21. 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, 107984-108011. <https://doi.org/10.48550/arXiv.2405.14458>
22. Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv*. <https://doi.org/10.48550/arXiv.2410.17725>

**Khrystyna Ruda**

PhD, Associate Professor of the Department of Information Security

Lviv Polytechnic National University, Lviv, Ukraine

ORCID:0000-0001-8644-411X

Khrystyna.S.Ruda@lpnu.ua

Vladyslav Kromkach

Assistant of the Department of Information Security

Lviv Polytechnic National University, Lviv, Ukraine

ORCID:0009-0001-5608-5715

vlad.kromkach@gmail.com

**SCENE-COMPLEXITY ADAPTIVE CASCADE FOR OBJECT DETECTION IN UAV IMAGERY:
WHEN SELECTIVE COMPUTE ESCALATION PAYS OFF**

Abstract. Unmanned aerial vehicles are a standard video-monitoring instrument in security tasks – from perimeter and critical-infrastructure protection to situational awareness – and detector ensembles raise the accuracy of object detection in aerial imagery at a manifold per-frame compute cost, which is critical for onboard systems. A Scene-Complexity Adaptive Cascade (SCAC) is proposed and investigated: a cheap first-tier detector processes every frame and computes a dimensionless, annotation-free complexity signal; only frames whose complexity exceeds a threshold are escalated to an expensive second-tier WBF ensemble. The threshold is not hand-tuned – it is calibrated by cross-fitting at the video-sequence level for a target average budget. The methodological contribution is an oracle feasibility gate: before the method is built, a greedy reference curve ordered by the realized per-frame ensemble gain is constructed from cached predictions, and realizable complexity signals are compared against it. The gate yields the negative part of the result: in the training domain (VisDrone-DET) the ensemble gain is dense across frames (75-90 % of frames improve), so realizable complexity signals do not beat uniform fusion at the same budget (the selective potential revealed only by a label-informed oracle remains out of their reach), and the popular uncertainty signal is consistently worse than random escalation. Under domain shift (zero-shot transfer to UAVDT with official condition labels), by contrast, the ensemble gain is sparse (41 % of frames; at night the ensemble hurts), and the cross-fitted cascade with a detection-density signal reaches the quality of the full ensemble (0.618 mAP@50, scene-balanced) at 45 % of its compute, significantly outperforming the single model (+0.022 on the full sample; bootstrap mean +0.017, 95 % CI [+0.002; +0.032]) and random escalation, including its scene-stratified variant (+0.019 on the full sample; bootstrap mean +0.016, 95 % CI [+0.001; +0.029]). The realized escalation profile (9 % night / 22 % day / 45 % fog) reproduces scene adaptivity without any scene classifier, bypassing the known bottleneck of routing – the unrecognizability of adverse conditions from frame appearance. A practical criterion is formulated: selective compute escalation pays off only where the expensive model's gain is sparse across frames – typically under domain shift, not in the training domain.

Keywords: object detection; UAV imagery; model cascade; detector ensemble; adaptive computation; domain shift; VisDrone; UAVDT.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 104117. <https://doi.org/10.1016/j.imavis.2021.104117>
2. Ruda, Kh. S., & Kromkach, V. O. (in press). Scene-dependent ensemble object detection in UAV aerial imagery: Limits of the utility of scene-based routing. *Scientific Bulletin of UNFU*.
3. Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 511-518). <https://doi.org/10.1109/CVPR.2001.990517>
4. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). BranchyNet: Fast inference via early exiting from deep neural networks. In *23rd International Conference on Pattern Recognition (ICPR)* (pp. 2464-2469). <https://doi.org/10.1109/ICPR.2016.7900006>



5. Kang, D., Emmons, J., Abuzaid, F., Bailis, P., & Zaharia, M. (2017). NoScope: Optimizing neural network queries over video at scale. *Proceedings of the VLDB Endowment*, 10(11), 1586-1597. <https://doi.org/10.14778/3137628.3137664>
6. Jiang, J., Ananthanarayanan, G., Bodik, P., Sen, S., & Stoica, I. (2018). Chameleon: Scalable adaptation of video analytics. In *Proceedings of the ACM SIGCOMM 2018 Conference* (pp. 253-266). <https://doi.org/10.1145/3230543.3230574>
7. Ghodrati, A., Bejnordi, B. E., & Habibian, A. (2021). FrameExit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 15608-15618). <https://doi.org/10.1109/CVPR46437.2021.01535>
8. Matsubara, Y., Levorato, M., & Restuccia, F. (2022). Split computing and early exiting for deep learning applications: Survey and research challenges. *ACM Computing Surveys*, 55(5), 1-30. <https://doi.org/10.1145/3527155>
9. 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>
10. Rahmath P., H., Srivastava, V., Chaurasia, K., Pacheco, R. G., & Couto, R. S. (2024). Early-exit deep neural network – A comprehensive survey. *ACM Computing Surveys*, 57(3), 1-37. <https://doi.org/10.1145/3698767>
11. 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>
12. 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. In *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.1701.06538>
13. 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2778-2788). <https://doi.org/10.1109/ICCVW54120.2021.00312>
14. 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>
15. Nazarkewych, M. A., & Oleksiv, N. T. (2024). Object recognition system based on the YOLO model. *Ukrainian Journal of Information Technology*, 6(1), 120-126. <https://doi.org/10.23939/ujit2024.01.120>
16. Filimonchuk, T. V., Koltun, Yu. M., & Maslov, M. K. (2026). Adaptive model for equipment recognition with attention mechanisms based on YOLOv11. *Information and Control Systems for Railway Transport*, 31(2). <https://doi.org/10.18664/iksz.v31i2.362205>
17. Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022). 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>
18. 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. In *Computer Vision – ECCV 2018* (pp. 370-386). https://doi.org/10.1007/978-3-030-01249-6_23
19. Saksena, S. K. (2025). *VisDrone detection model zoo* [Model collection]. Hugging Face. <https://huggingface.co/collections/dronefreak/visdrone-detection-model-zoo>
20. Wang, C.-Y., Yeh, I.-H., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. In *Computer Vision – ECCV 2024 (Lecture Notes in Computer Science, Vol. 15089)*, pp. 1–21. https://doi.org/10.1007/978-3-031-72751-1_1
21. 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, 107984-108011. <https://doi.org/10.48550/arXiv.2405.14458>
22. Khanam, R., & Hussain, M. (2024). YOLOv11: An overview of the key architectural enhancements. *arXiv*. <https://doi.org/10.48550/arXiv.2410.17725>

Отримано редакцією журналу / Received: 07.02.26

Прорецензовано / Revised: 20.02.26

Схвалено до друку / Accepted: 24.09.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.