УДК 004.93'1

DOI: https://doi.org/10.18664/ikszt.v31i2.362205

ФІЛІМОНЧУК Т. В., к.т.н, доцент,
КОЛТУН Ю.М., к.т.н, доцент,
МАСЛОВ М.К., магістр,
Харківський національний університет радіоелектроніки

Адаптивна модель розпізнавання техніки з механізмами уваги на базі YOLOv11

Анотація. Актуальність *дослідження зумовлена необхідністю підвищення ефективності систем комп'ютерного зору в умовах сучасної війни, де стрімкий розвиток безпілотних літальних апаратів і засобів розвідки потребує надійних інструментів для автоматизованого виявлення військової техніки в реальному часі. Існуючі детектори, попри високу швидкодію, демонструють зниження точності в разі роботи з малорозмірними, замаскованими об'єктами або в умовах складної фонової обстановки через втрату просторової інформації під час згорткових операцій. У зв'язку з цим удосконалення моделі архітектури нейронних мереж через інтеграцію механізмів уваги та оптимізацію обчислювальних процесів набуває особливої значущості.* ***Об'єктом дослідження*** *є процес автоматизованого виявлення та локалізації об'єктів військового призначення у відеопотоці оптичного діапазону.* ***Предметом дослідження*** *є методи структурної модифікації архітектури згорткової нейронної мережі YOLO з використанням механізмів координатної уваги та розділених детекторних модулів.* ***Результати.*** *У роботі запропоновано та реалізовано модифіковану модель архітектури нейромережі, головною особливістю якої є точкова інтеграція полегшеного модуля координатної уваги (Lite CA) перед шаром пірамідальної агрегації ознак, що дало змогу здійснити декомпозицію процесу просторового кодування, зберігши точну позиційну інформацію, яку зазвичай втрачають за стандартного глобального усереднення. Впровадження в модель складової «детекторна частина» забезпечило незалежну обробку задач класифікації та регресії координат. Експериментально підтверджено, що адаптивна модель досягає показника середньої точності mAP@0.5 на рівні 0.651, що на 8 % перевищує базовий рівень, і демонструє суттєве зростання повноти виявлення.* ***Висновки.*** *Запропонований підхід забезпечує ефективний баланс між точністю локалізації та обчислювальною складністю, гарантуючи високу стійкість детектора щодо візуальних перешкод та ефектів камуфляжу. Розроблена модель рекомендована для впровадження в автономні системи ситуаційної обізнаності та цілевказання, допомагаючи мінімізувати помилки пропуску цілі в бойових умовах.*

Ключові слова: *YOLOv11, CBAM, механізм уваги, детекція об'єктів, комп'ютерний зір, трансферне навчання, згорткові нейронні мережі, Python, PyTorch.*

Постановка проблеми

В умовах сучасного бойового простору важливим фактором успіху є досягнення інформаційної переваги та ситуаційної обізнаності. Візуальна інформація надходить із різнорідних джерел: бортових систем прицілювання та спостереження бронетехніки (танків, БМП, БТР), камер стаціонарних спостережних пунктів, а також безпілотних літальних апаратів. Масив відеоданих формує складне інформаційне середовище, ефективний аналіз якого людиною-оператором у режимі реального часу є ускладненим через фізичну втому, стрес і динаміку бойових дій.

На сьогодні стандартом для автоматизованої детекції об'єктів є нейромережеві архітектури сімейства YOLO (You Only Look Once), які завдяки високій швидкодії можуть бути розгорнуті на бортових обчислювачах мобільних платформ. Однак базові алгоритми YOLO, навчені на стандартних наборах даних, демонструють суттєве зниження ефективності в разі роботи в реальних умовах експлуатації військової техніки.

Основними проблемами, що виникають з обробкою відеопотоків із зазначених джерел, є такі:

- деградація візуального сигналу: відеоряд із камер бронетехніки або польових пунктів спостереження часто спотворений через вібрації під час руху, запиленість оптики, задимлення, складні погодні умови або низьку роздільну здатність сенсорів (тепловізорів, нічних прицілів). Стандартні згорткові мережі сприймають ці візуальні дефекти як шум, що призводить до пропуску цілей;
- варіативність ракурсів і масштабів: система має однаково ефективно розпізнавати техніку як у фронтальній проєкції (вигляд з екрана навідника танка/БТР на рівні землі), так і верхній проєкції (вигляд з БПЛА). Базові моделі часто мають обмежену здатність генералізації таких різнорідних просторових ознак;
- маскування та часткове перекриття (Occlusion): в умовах наземного бою техніка противника часто використовує природні укриття (лісосмуги, складки рельєфу) або засоби маскування (сітки, «накидки»).

Локальні рецептивні поля стандартних CNN не дають змогу ефективно виокремити об'єкт, якщо видима лише його незначна частина (наприклад лише башта або гусениці).

Наукова задача полягає в модифікації архітектури нейронної мережі YOLO через інтеграцію механізмів уваги (Attention Mechanisms), зокрема модуля CBAM, що допоможе створити адаптивну систему, здатну динамічно перерозподіляти «увагу» мережі на найбільш інформативні ділянки кадру, ігноруючи перешкоди та фон. Таке рішення забезпечить підвищення точності та надійності автоматизованого цілевказання як для екіпажів бойових машин, так і операторів пунктів спостереження, незалежно від типу відеосенсора.

Аналіз останніх досліджень і публікацій

Для сучасного етапу розвитку систем комп'ютерного зору характерне активне впровадження нейромережевих архітектур сімейства YOLO, які стали стандартом де-факто для задач детекції об'єктів у реальному часі. Аналіз наукової періодики свідчить про зміщення фокусу досліджень від простого підвищення точності до оптимізації архітектур для роботи в складних умовах (обмежені обчислювальні ресурси, малі об'єкти, перешкоди) і глибокої інтеграції додаткових механізмів уваги.

У фундаментальній оглядовій роботі [1] проведено глибокий системний аналіз еволюції архітектур YOLO від першої версії до найновіших ітерацій YOLOv8 і YOLO-NAS. Наукова новизна роботи полягає в детальному розборі переходу від якірних (anchor-based) до безякірних (anchor-free) підходів, що дало змогу значно спростити процес навчання та підвищити генералізацію моделі. Автори детально розглядають структурні зміни в блоках Backbone, зокрема заміну модулів C3 на більш градієнтно-ефективні C2f, а також повну реструктуризацію елемента «Голова» (Head) на відокремлені гілки (decoupled head) для класифікації та регресії координат. Особливу увагу приділено проблемі компромісу між швидкістю інференсу і точністю (mAP), де новітні моделі демонструють значну перевагу завдяки використанню методів автоматичного пошуку нейронної архітектури (Neural Architecture Search, NAS).

У роботі [2] комплексно досліджено проблему оптимізації моделей YOLOv8 для задач розпізнавання специфічних цілей в умовах обмежених вибірок даних. Ноу-хау авторів полягає в розробленні адаптивної стратегії навчання, що враховує залежність ефективності детекції від роздільної здатності вхідного зображення та масштабу цільових об'єктів. Автори провели серію експериментів із різними розмірностями моделі (nano, small, medium) і довели, що просте збільшення глибини мережі не завжди призводить до покращення результатів на малих об'єктах. Натомість критично важливим фактором визначено етап попередньої обробки даних (Data Preprocessing) і використання просунутих технік аугментації, таких як Mosaic і Mixup, що допомагає моделі «бачити» об'єкти в нових контекстах. Результати підтверджують, що правильне налаштування гіперпараметрів (зокрема learning rate scheduler) дає змогу легким моделям досягати точності, порівнянної з важкими архітектурами, але за значно менших обчислювальних витрат.

У дослідженні [3] запропоновано практичну реалізацію end-to-end системи розпізнавання об'єктів на базі YOLO, акцентуючи на безшовній інтеграції нейромережі з апаратними модулями відеоспостереження. Основну увагу приділено технічним аспектам розгортання моделі (Deployment) у реальних умовах експлуатації, де критичними факторами є динамічні зміни освітлення, погодні умови та складні ракурси зйомки. Автори проаналізували вплив різних форматів вхідного відеопотоку (RTSP, HTTP) на затримку (latency) обробки та запропонували алгоритм буферизації кадрів, що мінімізує пропуски детекції за різких рухів камери. Експериментально доведено, що використання оптимізованих тензорних ядер GPU у поєднанні з архітектурою YOLO допомагає досягти стабільного фреймрейту (FPS) вище 30 кадрів за секунду навіть для обробки зображень високої чіткості, що є критичним показником для систем безпеки та моніторингу периметра.

Окремий напрям досліджень, висвітлений у роботі [4], стосується складного завдання ідентифікації та супроводу рухомих об'єктів (Object Tracking). Автори пропонують комплексний підхід щодо побудови системи, яка не лише детектує об'єкти в кожному кадрі, але й відновлює їхні траєкторії руху, розв’язуючи проблему перекриття (occlusion). У статті детально проаналізовано ефективність інтеграції детектора YOLO з алгоритмами трекінгу, такими як SORT і DeepSORT. Основним висновком із роботи є те, що стабільність трекінгу безпосередньо корелює з якістю роботи детектора: навіть незначні помилки в локалізації (bounding box) призводять до збоїв в ідентифікації (ID switch). Автори демонструють, як модифікація згорткових шарів для кращого виділення просторових ознак допомагає зменшити кількість помилкових спрацьовувань на складних фонах, підвищуючи загальну надійність системи моніторингу.

У роботі [5] розглянуто вузькоспеціалізоване застосування нейромережевих технологій для виявлення небезпечних об'єктів (зокрема вибухонебезпечних предметів) у вкрай складних візуальних середовищах, включаючи підводну зйомку. Автор досліджує методи адаптації архітектур YOLO для роботи із зображеннями низької якості з низьким контрастом, розмиттям і специфічними кольоровими спотвореннями. Запропонована методологія включає використання трансферного навчання (Transfer Learning), коли модель попередньо тренують на великих загальних датасетах, а потім донавчають (fine-tuning) на специфічних даних. Ноу-

хау полягає у використанні спеціальних фільтрів попередньої обробки зображень для підкреслення країв об'єктів перед подаванням їх у нейромережу. Цей підхід є релевантним для будь-яких задач, де об'єкти мають складну форму або частково замасковані середовищем.

У роботі [6] наведено альтернативний підхід щодо детекції, базований на архітектурі трансформерів (Vision Transformers), а саме RT-DETR (Real-Time DEtection TRansformer). На відміну від згорткових мереж, таких як YOLO, що аналізують локальні ознаки, трансформери використовують механізм глобальної уваги (Self-Attention) для розуміння контексту всього зображення. Автори порівнюють ефективність YOLOv8 і RT-DETR і зазначають, що трансформерні моделі краще справляються з ситуаціями сильного перекриття об'єктів і відсутності чітких візуальних ознак. Хоча трансформери потребують більше ресурсів для навчання, їхня здатність моделювати довгі залежності між пікселями підкреслює критичну важливість використання механізмів уваги для досягнення високої точності детекції в складних умовах.

У роботі [7] запропоновано концептуальну модель системи управління процесами розпізнавання образів, яка базована на інтеграції сучасних методів штучного інтелекту. Автори фокусують увагу на проблемі побудови адаптивних систем, здатних ефективно функціонувати в умовах динамічного вхідного потоку даних. Наукова цінність дослідження полягає в розробленні структурної схеми, яка систематизує етапи життєвого циклу моделі: від попередньої обробки даних до валідації результатів. Автори зазначають, що просте використання потужної архітектури (CNN) без належної організації конвеєра обробки даних та управління параметрами не гарантує стабільності результату. Цей підхід корелює з необхідністю оптимізації архітектури типу YOLO для специфічних задач, підтверджуючи важливість комплексного підходу щодо проєктування систем комп'ютерного зору, де алгоритми розпізнавання тісно пов'язані з модулями ухвалення рішень.

Узагальнюючи проведений аналіз наукових джерел [1-7], можна зробити висновок, що існуючі базові моделі детекції (зокрема стандартні версії YOLO) забезпечують високу швидкодію, проте мають суттєві обмеження для роботи в складних умовах експлуатації. Основними недоліками є втрата інформації про малорозмірні об'єкти в процесі згорткових перетворень, недостатня стійкість щодо візуальних перешкод (маскування, погане освітлення) і відсутність механізмів глобального фокусування уваги, притаманних трансформерним архітектурам.

Більшість розглянутих рішень сфокусовано на окремих аспектах: або попередній обробці даних, або архітектурних змінах, або оптимізації інференсу. Проте для досягнення високої точності та надійності в системах ситуаційної обізнаності необхідний комплексний підхід.

На основі цього виникає необхідність розроблення вдосконаленої моделі, яка б поєднувала ефективність згорткових мереж з адаптивністю механізмів уваги. Формально узагальнену модель автоматизованої системи детекції MASD можна подати як впорядкований кортеж функціональних компонентів:

$$M_{ASD} = \left(PreP, FE, FPN, DH, LFn, InfE \right), \qquad (1)$$

де PreP (Preprocessing module) – модуль базової попередньої обробки;

FE (Feature Extractor) – стандартний екстрактор ознак Backbone;

FPN (Feature Pyramid Network) – класичний модуль агрегації ознак без механізмів уваги;

DH (Detection Head) – детекторна частина;

LFn (Loss Function) – функція втрат;

InfE (Inference Engine) – підсистема інференсу.

Зазначена формалізація дає змогу чітко визначити напрям подальшого дослідження як оптимізацію кожного з наведених компонентів для досягнення ефекту з точності та швидкодії системи.

Мета дослідження

Метою дослідження є підвищення ефективності автоматизованого виявлення та класифікації об’єктів військової техніки у відеопотоці реального часу через структурну модифікацію архітектури згорткової нейронної мережі YOLO. Досягнення мети забезпечено через інтеграцію в модуль агрегації ознак механізму координатної уваги та впровадження механізму незалежної обробки задач класифікації та локалізації, що в поєднанні зі стохастичними методами аугментації допомагає системі адаптивно фокусуватися на інформативних ділянках зображення, гарантуючи стійку селекцію малорозмірних і замаскованих цілей в умовах складної фонової обстановки.

Викладення основного матеріалу

Сучасна парадигма ведення бойових дій зазнає фундаментальних змін із впливом тотальної цифровізації та роботизації поля бою. Важливим фактором, що визначає тактичну перевагу, стає не лише вогнева міць, а і швидкість циклу ухвалення рішень (OODA loop), яка безпосередньо залежить від ефективності розвідувальних систем. У цьому контексті критичного значення набувають технології комп’ютерного зору, інтегровані в безпілотні авіаційні комплекси та системи управління вогнем, оскільки вони дають змогу автоматизувати процес виявлення та ідентифікації цілей без безпосередньої участі оператора, мінімізуючи людський фактор і час реакції.

На сьогодні стандартом у задачах детекції об'єктів реального часу є використання глибоких згорткових нейронних мереж (CNN), зокрема

одностадійних детекторів архітектури YOLO. Їх широке розповсюдження зумовлене архітектурною здатністю виконувати передбачення класів і координат об'єктів за один прохід мережі, що забезпечує високу кадрову частоту обробки навіть на вбудованих обчислювальних платформах з обмеженим енергоспоживанням (Edge AI). Останні ітерації цієї архітектури, такі як YOLO, демонструють вражаючі результати на стандартних еталонних наборах даних, використовуючи вдосконалені методи агрегації ознак та оптимізовані функції активації.

Однак, незважаючи на загальну ефективність, застосування базових архітектур CNN у специфічних умовах військового моніторингу виявляє низку критичних недоліків. Фундаментальна проблема полягає в природі операції згортки, яка оперує в межах обмеженого локального рецептивного поля. Із проходженням сигналу через глибокі шари мережі та операції зниження дискретизації (downsampling) просторову інформацію про малорозмірні об'єкти часто втрачають або вона «розмита» на фоні високочастотного шуму.

Ситуація значно ускладнена специфікою вхідних даних у бойових умовах. Військова техніка часто має низький візуальний контраст відносно навколишнього середовища, використовує камуфляжне фарбування, маскувальні сітки або знаходиться в умовах природного затінення. Крім того, фактори навколишнього середовища, такі як атмосферні опади, туман, задимленість або низька освітленість, створюють складні візуальні патерни, які стандартна згорткова мережа може інтерпретувати як частину ландшафту. У таких умовах стандартні механізми вилучення ознак є недостатньо селективними, що призводить до зростання помилок першого (пропуск цілі) і другого (хибне спрацювання) роду.

Традиційні методи розв'язання цієї проблеми простим збільшенням глибини або ширини нейронної мережі є неприйнятними для систем військового призначення, оскільки це призводить до експоненційного зростання обчислювальної складності та затримок інференсу, що є критичним для задач реального часу. Отже, виникає необхідність пошуку архітектурних рішень, які б допомогли підвищити чутливість мережі до значущих ознак без суттєвого збільшення обчислювального навантаження.

Перспективним напрямом подолання описаних обмежень є інтеграція в архітектуру детектора механізмів уваги (Attention Mechanisms). На відміну від стандартних згорток, які обробляють усі пікселі зображення з однаковим пріоритетом, механізми уваги дають змогу мережі динамічно перерозподіляти ваги, фокусуючись на найбільш інформативних просторових зонах або канальних ознаках. Зокрема, механізми, подібні до інтегрованого блока координатної уваги (Coordinate Attention), здатні моделювати довгострокові залежності та кодувати точну позиційну інформацію, яку зазвичай втрачають із використанням глобального усереднення.

Для розв'язання поставленої науково-технічної задачі подано модифіковану модель автоматизованої системи детекції, яка інтегрує класичні згорткові методи вилучення ознак із новітніми механізмами просторової та канальної уваги. Запропонована архітектура базована на платформі YOLOv11, суттєво вдосконаленій через впровадження блоків Coordinate Attention та оптимізацію конвеєра попередньої обробки даних. Такий підхід допомагає подолати обмеження стандартних детекторів для роботи з малорозмірними та замаскованими об'єктами в складних умовах спостереження.

Формально запропоновану модель системи детекції об'єктів (ODS) можна описати як упорядкований кортеж функціональних компонентів:

$$\mathrm{ODS} = \left\{ \mathrm{PreP}, \mathrm{FE}, \mathrm{FPN}_{\mathrm{CA}}, \mathrm{DH}, \mathrm{LFn}, \mathrm{InfE} \right\}, \quad 2)$$

де PreP (Preprocessing module) – модуль попередньої обробки та аугментації даних;

FE (Feature Extractor) – модуль вилучення ознак (Backbone);

$\mathrm{FPN}_{\mathrm{CA}}$ (Feature Pyramid Network with attention) – модифікований модуль агрегації ознак;

DH (Detection Head) – детекторна частина;

LFn (Loss Function mechanism) – механізм обчислення функції втрат;

InfE (Inference Engine) – програмний рушій виконання моделі.

Структурна унікальність запропонованого рішення базована на комплексному вдосконаленні базових підходів. По-перше, модуль попередньої обробки (PreP) розширено алгоритмами стохастичної трансформації, що забезпечує інваріантність детектора до геометричних спотворень і змін освітлення. По-друге, архітектура нейронної мережі модифікована через інтеграцію механізму CA у модуль агрегації ознак $\mathrm{FPN}_{\mathrm{CA}}$, що дає змогу системі враховувати глобальний контекст і зберігати точну позиційну інформацію, яка зазвичай нівельована під час зниження просторової розмірності. По-третє, для забезпечення роботи в реальному часі розроблено InfE, яка оптимізує взаємодію з апаратним забезпеченням і стабілізує відеопотік.

Порівняно з традиційними підходами запропонована модель забезпечує вищу точність локалізації об'єктів завдяки адаптивному фокусуванню уваги, зберігаючи при цьому високу швидкодію, характерну для одностадійних детекторів. Використання принципів трансферного навчання дає змогу ефективно адаптувати систему до специфічних класів об'єктів навіть за умови обмеженого обсягу навчальних даних. Усі модулі системи реалізовані мовою Python із використанням бібліотек глибокого навчання, що забезпечує гнучкість і масштабованість рішення.

Функціонування наведеної системи детекції розпочинається з роботи модуля PreP, який реалізує комплексний підхід щодо підготовки вхідних даних. Формально структуру модуля попередньої обробки можна подати як сукупність алгоритмів трансформації даних:

$$\mathrm{PreP} = \left\{ \mathrm{MDA}, \mathrm{MixUp}, \mathrm{Aug}_{\mathrm{HSV}}, \mathrm{Aug}_{\mathrm{Geom}} \right\}, \quad (3)$$

де MDA (Mosaic Data Augmentation) – алгоритм мозаїчного суміщення кадрів;

MixUp – метод лінійного змішування зображень;

$\mathrm{Aug}_{\mathrm{HSV}}$ – фотометричні спотворення в колірному просторі;

$\mathrm{Aug}_{\mathrm{Geom}}$ – геометричні афінні трансформації.

В умовах реального бойового застосування, де відеопотік надходить із камер безпілотних літальних апаратів (БПЛА) або наземних роботизованих платформ, вхідні зображення мають високу варіативність масштабу, ракурсу та умов освітлення. У таких сценаріях класичний підхід щодо попередньої обробки, що обмежений детермінованою зміною розміру (resize) до вхідної роздільної здатності мережі (наприклад 640 на 640 пікселів), виявляється недостатнім. Просте масштабування часто призводить до втрати дрібних деталей, критично важливих для розпізнавання малорозмірних об'єктів, і не забезпечує необхідну інваріантність моделі до геометричних спотворень.

Для розв'язання цієї проблеми в модулі PreP реалізовано концепцію стохастичної трансформації, яка має на меті штучне розширення простору ознак навчальної вибірки. Реалізація запропонованої методики дає змогу наблизити розподіл даних, на якому навчається модель, до реального розподілу візуальних сцен на полі бою.

Важливим елементом стратегії аугментації є алгоритм мозаїчного суміщення (MDA). На відміну від традиційних методів, які оперують поодинокими зображеннями, MDA формує один навчальний зразок через синтез чотирьох випадкових зображень із датасету. Процес генерації мозаїки складається з таких етапів:

- вибірка та ініціалізація: вибирають індекси чотирьох випадкових зображень I_1, I_2, I_3, I_4;
- генерують цільове полотно розміром W×H;
- визначення точки центрування: генерують випадкову точку перетину (x_c, y_c) у межах заданого діапазону координат, яка стає центром стикування чотирьох зображень;
- компонування та обрізка: кожне з чотирьох зображень розміщено у відповідному квадранті відносно точки (x_c, y_c). Частини зображень, що виходять за межі цільового вікна розміром W×H навколо центру, відсікають.

Математична логіка функціонування алгоритму MDA передбачає перерахунок координат обмежувальних рамок (Bounding Boxes) для кожного об'єкта відносно нових меж синтезованого зображення. Після зміщення фрагмента на вектор dx, dy нові координати (x', y') обчислюють з урахуванням кліпінгу (обрізання) за шириною W і висотою H полотна:

$$B = \begin{cases} x'_{\min} = \max\left(0, x_{\min} + dx\right) \\ y'_{\min} = \max\left(0, y_{\max} + dy\right) \\ x'_{\max} = \min\left(W, x_{\max} + dx\right) \\ y'_{\max} = \min\left(H, y_{\max} + dy\right) \end{cases}, \quad (4)$$

де $x'_{\min}$, $y'_{\min}$, $x'_{\max}$, $y'_{\max}$ – координати вершин обмежувальної рамки на новому полотні мозаїки;

$x_{\min}$, $y_{\min}$, $x_{\max}$, $y_{\max}$ – початкові координати рамки об'єкта на вихідному зображенні;

dx, dy – вектори зміщення зображення відносно центра мозаїки;

W, H – ширина та висота цільового зображення для навчання.

Якщо площа нової рамки стає меншою за порогове значення, об'єкт виключають із розмітки.

Використання алгоритму MDA дає системі суттєві переваги: по-перше, дає змогу моделі ефективно навчатися розпізнавати об'єкти, що знаходяться на периферії кадру (частково обрізані), що є типовим для відеопотоку з рухомого дрона. По-друге, через значне варіювання масштабу з компонуванням мережа стає більш чутливою до малорозмірних цілей. По-третє, завдяки наявності об'єктів із чотирьох різних сцен в одному тензорі стабілізовано обчислення статистик Batch Normalization, що допомагає ефективно навчати модель навіть за малого розміру батчу (Mini-batch size).

Для підвищення завадостійкості моделі щодо візуальних перешкод та ефекту накладання об'єктів застосовують алгоритм MixUp, в основі якого лежить принцип навчання на віртуальних прикладах, отриманих опуклою комбінацією пар зображень і їхніх міток.

Формально для двох випадкових пар «зображення-мітка» новий навчальний приклад tilde_x, tilde_y розраховують за формулами лінійної інтерполяції:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda) x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda) y_{Ij} \end{aligned},$$

де λ [0, 1] – коефіцієнт змішування.

У контексті детекції військової техніки візуальний результат роботи алгоритму MixUp нагадує напівпрозоре накладання одного об'єкта на інший або на фоновий ландшафт, що дає змогу емулювати реальні фізичні явища:

- сприйняття об'єктів через середовища, що розсіюють світло (туман, дим, пилова завіса);

- ефект часткової прозорості маскувальних сіток;

- візуальне злиття камуфльованого об'єкта зі складним фоном (лісосмуга, міська забудова).

Регуляризаційний ефект MixUp запобігає перенавчанню мережі та «запам'ятовуванню» чітких контурів, змушуючи модель покладатися на більш робастні ознаки.

Фінальним етапом у конвеєрі PreP є застосування фотометричних Aug_{HSV} і геометричних Aug_{Geom} спотворень. Колірна корекція відбувається у просторі HSV (Hue, Saturation, Value), оскільки він краще відповідає людському сприйняттю кольору, ніж RGB. Для кожного зображення виконано стохастичний зсув каналів:

$$\begin{aligned} H_{new} &= \left(H + \delta_h\right) \bmod 180 \\ S_{new} &= \text{clip}\left(S \cdot \delta_s\right) \\ V_{new} &= \text{clip}\left(V \cdot \delta_v\right) \end{aligned}, \quad (6)$$

де H_{new}, S_{new}, V_{new} – модифіковані значення відтінку, насиченості та яскравості пікселя відповідно;

H, S, V – початкові значення компонентів у колірному просторі HSV;

δ_h, δ_s, δ_v – випадкові коефіцієнти мультиплікативного шуму для кожного каналу;

clip() – функція обмеження значень у допустимому діапазоні [0, 255].

Застосування колірних корекцій відтворює варіативність умов освітлення (яскраве сонце, хмарність, сутінки) і спектральні особливості камер. Геометричні трансформації Aug_{Geom} включають випадкові повороти, віддзеркалення та перспективні викривлення, гарантуючи інваріантність детектора до положення камери відносно горизонту.

Після попередньої обробки вхідний тензор надходить до модуля вилучення ознак (FE), який виконує функцію Backbone нейронної мережі. Архітектурно цей модуль визначено як композицію таких функціональних блоків:

$$FE = \left\{\text{CSPNet}, \text{C2f}, \text{SPPF}\right\}, \quad (7)$$

де CSPNet (Cross Stage Partial Network) – базова концепція мережі з частковими перехресними стадіями;

C2f (Cross Stage Partial Bottleneck with 2 convolutions) – модифікований блок агрегації градієнтів;

SPPF – модуль швидкої пірамідальної агрегації просторових ознак.

Архітектурна парадигма модуля FE базована на концепції мережі з частковими перехресними стадіями (CSPNet), яка була спеціально розроблена для розв'язання проблеми надлишковості градієнтної інформації в глибоких згорткових мережах.

Фундаментальна ідея CSPNet полягає в розділенні карти ознак базового шару на дві частини. Одна частина проходить через щільний блок (Dense Block) або серію згорткових перетворень, а інша частина передана транзитом і конкатенована з виходом щільного блока на наступній стадії переходу. Такий підхід допомагає досягти багатшого градієнтного потоку, оскільки градієнти від помилки поширюються різними шляхами, що запобігає їх дублюванню під час оновлення ваг. Реалізація такого підходу сприяє зменшенню обчислювальної складності моделі зі збереженням або навіть підвищенням точності детекції, що є критичним для роботи на пристроях з обмеженими ресурсами.

В архітектурі YOLO класичний модуль C3, що використаний у попередніх версіях, замінено на вдосконалений блок C2f. Основна відмінність полягає у структурі розгалуження потоків і кількості вихідних з'єднань, що дає змогу моделі захоплювати більш багату семантичну інформацію.

Процес обробки інформації всередині блока C2f можна формалізувати так: вхідний тензор X_{in} спочатку проходить через згортку $Conv_1$, яка зменшує кількість каналів, далі отриманий тензор розділяється на дві гілки:

$$Y_1, Y_2 = \text{Split}\left(\text{Conv}_1\left(X_{in}\right)\right), \quad (8)$$

де X_{in} – вхідний тензор ознак, що надходить у блок C2f;

$Conv_1$ – згортка 1×1, яка зменшує розмірність каналів вхідного тензора;

Split – операція розділення тензора на дві рівні частини вздовж канального виміру;

Y_1 – частина потоку, що передана транзитом (skip connection) для збереження низькорівневих деталей;

Y_2 – частина потоку, що спрямована на глибоку обробку через серію Bottleneck-шарів.

Кожен блок Bottleneck складається з двох згорток 3×3 і використовує залишкові зв'язки (Residual connections), які відіграють головну роль у збереженні амплітуди градієнта з проходженням через глибокі шари, запобігаючи проблемі згасання градієнта (Vanishing Gradient Problem).

Унікальною особливістю C2f є те, що він агрегує не тільки вихід останнього Bottleneck-блока, а і проміжні виходи всіх внутрішніх блоків. Усі ці потоки об'єднані операцією конкатенації та проходять через фінальну згортку $Conv_2$ для відновлення вихідної розмірності каналів:

$$X_{out} = \text{Conv}_2\left(\text{Concat}\left(Y_1, Y_2, B_1\left(Y_2\right), B_2\left(B_1\left(\right.\right.\right.\right. \quad (9$$

де X_{out} – вихідний тензор блока C2f;

$B_i(Y_2)$ – вихід і-го послідовного блока Bottleneck;

Concat – операція конкатенації (об'єднання) усіх проміжних потоків ознак, що збагачує семантичне подання;

$Conv_2$ – фінальна згортка 1×1, що інтегрує об'єднані ознаки та відновлює необхідну кількість каналів.

Запропонована архітектура забезпечує більш ефективне злиття низькорівневих і високорівневих ознак, що підвищує здатність мережі розрізняти текстури та дрібні деталі об'єктів.

Завершальним елементом екстрактора ознак є модуль SPPF (Spatial Pyramid Pooling Fast), який призначений для розширення рецептивного поля мережі та агрегації контекстної інформації з різних масштабів. На відміну від класичного SPP, який використовує паралельні гілки субдискретизації з різними ядрами, SPPF реалізує послідовну каскадну схему, що є більш обчислювально ефективним рішенням.

Алгоритм роботи SPPF полягає в послідовному застосуванні операції максимізації (Max Pooling) із фіксованим розміром ядра 5×5 і кроком 1 (із використанням padding для збереження просторової розмірності). Нехай вхідний тензор модуля – F_0. Тоді послідовність перетворень має вигляд

$$\begin{aligned} F_1 &= \mathrm{MaxPool}_{5\times5}\left(F_0\right) \\ F_2 &= \mathrm{MaxPool}_{5\times5}\left(F_1\right), \\ F_3 &= \mathrm{MaxPool}_{5\times5}\left(F_2\right) \end{aligned} \tag{10}$$

де F_0 – вхідний тензор ознак, що надходить у модуль SPPF (вихід останнього блока C2f Backbone);

F_1, F_2, F_3 – послідовні тензори ознак, отримані після застосування операції Max Pooling;

$MaxPool_{5\times5}$ – операція вибору максимального значення з вікна 5×5 пікселів.

Оскільки послідовне виконання двох таких процедур субдискретизації математично еквівалентне охопленню рецептивного поля 9×9, а трьох – 13×13, запропонована структура уможливлює мультимасштабний аналіз ознак без залучення фільтрів великої розмірності, які суттєво уповільнюють обчислення. Результуючий тензор модуля сформований об'єднанням (конкатенацією) початкового масиву даних із результатами всіх етапів обробки, після чого застосовано фінальну згортку 1×1:

$$X_{SPPF} = \mathrm{Conv}\left(\mathrm{Concat}\left(F_0, F_1, F_2, F_3\right)\right), \tag{11}$$

де X_{SPFF} – вхідний тензор у модуль SPPF;

Concat – об'єднання виходів послідовних просторових ознак, що допомагає мережі «бачити» об'єкти різного масштабу одночасно;

Conv – згортковий шар, що змішує агреговані ознаки.

Завдяки такій інтеграції мережа отримує здатність ефективно поєднувати локальні ознаки об'єкта з його глобальним контекстом, що є критично важливим для коректної класифікації об'єктів на складному фоні.

Модуль агрегації ознак FPN_{CA} у запропонованій моделі є модифікацією класичної схеми і структурно описана як

$$\mathrm{FPN}_{CA} = \left\{\mathrm{PANet}, \mathrm{CA}\right\}, \tag{12}$$

де PANet (Path Aggregation Network) – мережа агрегації шляхів для об'єднання ознак різних масштабів;

CA (Coordinate Attention) – інтегрований блок координатної уваги, що забезпечує просторову селекцію.

У класичній архітектурі одностадійних детекторів (YOLOv5/v8/v11) модуль агрегації ознак (Neck) зазвичай реалізований на базі схеми PANet. Хоча цей підхід ефективно комбінує семантичну інформацію з глибоких шарів і деталізацію з поверхневих, він має суттєвий недолік для роботи зі складними сценами: відсутність явного механізму фокусування уваги. Стандартні згорткові операції обробляють усі пікселі в межах рецептивного поля з однаковим пріоритетом, що призводить до «забруднення» корисного сигналу фоновим шумом, особливо в умовах камуфляжу.

Спроби інтеграції класичних механізмів уваги, таких як SE (Squeeze-and-Excitation) або CBAM (Convolutional Block Attention Module), часто не дають бажаного приросту точності в задачах локалізації об'єктів. Основна причина полягає у використанні ними операції глобального усереднення GAP (Global Average Pooling) для стиснення просторової інформації в одновимірний вектор каналів.

Із застосуванням методу глобального усереднення до вхідного масиву ознак відбувається перетворення двовимірних карт активації в одновимірний вектор, де значення кожного каналу сформовано обчисленням середнього арифметичного всіх пікселів. Таке стиснення даних призводить до повної втрати інформації про просторову структуру зображення, оскільки нівельована прив'язка ознак до конкретних координат. Якщо для задач звичайної класифікації це прийнятно, оскільки головною метою є лише ідентифікація наявності об'єкта, то для задач детекції, де критично важливою є точна локалізація меж, такий підхід фактично знищує позиційну інформацію, знижуючи чутливість моделі до розташування цілі.

Для подолання цього обмеження в архітектуру FPN_{CA} інтегровано механізм CA, який дає змогу моделювати залежності між каналами, зберігаючи при цьому точну позиційну інформацію.

Головною ідеєю методу є декомпозиція (розкладання) процесу просторового кодування. Замість традиційного підходу, що передбачає миттєве згортання всієї двовимірної карти ознак у єдине число, алгоритм виконує дві паралельні операції агрегації вздовж координатних осей. Це допомагає окремо обробляти інформацію для вертикального та горизонтального напрямків. Отже, для вхідного тензора X накопичення ознак відбувається незалежно за висотою H і шириною W, а результат цієї трансформації для c-го каналу на висоті H визначено виразом

$$z_c^h(h) = \frac{1}{W}\sum_{0\le i<W} x_c(h,i), \qquad (13)$$

де $z_c^h(h)$ – значення глобально-усередненої ознаки для c-го каналу на висоті H;

$x_c(h, i)$ – значення пікселя вхідної карти ознак;

W – ширина карти ознак, вздовж якої відбувається усереднення.

Аналогічно результат агрегації для c-го каналу на ширині W визначено як

$$z_c^w(w) = \frac{1}{H}\sum_{0\le j<H} x_c(j,w), \qquad (14)$$

де H – висота карти ознак.

У результаті виконання цих операцій сформовано два незалежні набори агрегованих ознак: один відповідає за вертикальну складову зображення, а інший за горизонтальну. Такий підхід дає модулю змогу одночасно враховувати глобальний контекст уздовж однієї просторової осі та зберігати точну інформацію про позицію об'єкта вздовж іншої. Надалі отримані дескриптори піддають процедурі кодування для моделювання взаємозв'язків між каналами, під час якої відбувається їхнє просторове перетворення та подальше об'єднання в єдину структуру для спільної обробки.

Сформований тензор поданий на вхід згорткового шару $1\times1(Conv_1)$, який зменшує кількість каналів відповідно до коефіцієнта редукції r (у реалізації використано r = 32), що дає змогу зменшити обчислювальну складність. До результату згортки застосовано нелінійну функцію активації $H_{swish}(\delta)$, яка забезпечує гладкість градієнтів:

$$f = \delta\left(Conv_1\left(\left[z^h, z^w\right]\right)\right), \qquad (15)$$

де f – проміжний тензор ознак зі зменшеною розмірністю каналів;

$[z^h, z^w]$ – операція конкатенації векторів висоти і ширини;

$Conv_1$ – згортка 1×1, що моделює взаємозв'язки між каналами та зменшує їхню кількість на коефіцієнт редукції r;

δ – нелінійна функція активації (Hard Swish), що забезпечує кращу виразність ознак.

На наступному етапі проміжний тензор розділяється на два незалежні потоки даних, що відповідають вертикальному та горизонтальному просторовим вимірам. Після цього до кожного з отриманих масивів застосовано окрему операцію згортки, завданням якої є відновлення початкової глибини каналів до значення, що було на вході блока. Фінальні карти уваги генеровані через застосування сигмоїдальної функції активації σ:

$$g^h = \sigma\left(Conv_h\left(f^h\right)\right),$$
$$g^w = \sigma\left(Conv_w\left(f^w\right)\right), \qquad (16)$$

де g^h, g^w – фінальні вектори ваг уваги для висоти і ширини;

f^h, f^w – розділені частини проміжного тензора f;

$Conv_h$, $Conv_w$ – згорткові шари, що відновлюють початкову кількість каналів C;

σ – сигмоїдальна функція активації, що перетворює значення в діапазон [0, 1], інтерпретуючи їх як імовірність важливості ознаки.

Результуючий вихід блока CA обчислюють як поелементний добуток вхідного тензора X на отримані карти уваги. Карти уваги автоматично розтягуються до розмірності вхідного тензора:

$$Y_c(i,j) = x_c(i,j)\times g_c^h(i)\times g_c^w(j), \qquad (17)$$

де Y – вихідний тензор модуля CA;

x – вхідний тензор ознак;

g^h, g^w – вектори уваги, які діють як просторовий фільтр, підсилюючи сигнал у зонах розташування цільових об'єктів і пригнічуючи фоновий шум.

Фінальна формула (17) демонструє суть методу: значення кожного пікселя $x_c(i,j)$ модульована двома коефіцієнтами: $g_c^h(i)$ визначає важливість рядка (i) для каналу (c), а $g_c^w(j)$ – важливість стовпця (j). Отже, нейронна мережа отримує можливість адаптивно підсилювати сигнал у зонах, де ймовірно знаходиться цільовий об'єкт (танки, бронетехніка) і пригнічувати сигнал у зонах фону (небо, трава), що значно підвищує відношення сигнал/шум перед подаванням даних на детекторну частину.

Завершальним етапом обробки інформації в нейромережевій архітектурі є функціонування модуля детекторної голови (DH). Концептуально архітектуру цього модуля можна подати як композицію трьох структурних принципів:

$$DH = \{Decoupled, AnchorFree, DFL\}, \quad (18)$$

де Decoupled – архітектура з розв'язаними гілками класифікації та регресії для усунення конфлікту ознак;

AnchorFree – без'якірний метод передбачення координат, що спрощує навчання;

DFL (Distribution Focal Loss) – механізм фокальних втрат розподілу для підвищення точності локалізації.

У розробленій моделі першочергово імплементовано концепцію «розв'язаної голови» Decoupled, яка розв'язує фундаментальну проблему конфлікту ознак (Feature Misalignment), притаманну класичним одностадійним детекторам. Суть проблеми полягає в тому, що задачі класифікації та локалізації потребують від нейронної мережі фокусування на різних типах інформації: для коректного визначення класу об'єкта критично важливими є інваріантні семантичні ознаки (текстура, форма), тоді як для точної регресії координат необхідна детальна просторова інформація про межі та градієнти контурів.

Архітектурно принцип Decoupled Head реалізовано через розділення вхідного потоку даних на дві незалежні паралельні гілки обробки. Формально структуру цього модуля для кожного рівня масштабу піраміди ознак P_i можна подати як композицію функцій

$$Pi = \{\phi_{cls}(Conv(P_i)), \phi_{reg}(Conv(P_i))\}, \quad (19)$$

де Conv – згортковий шар для зменшення розмірності каналів;

φ_{cls} – послідовність згорткових операцій гілки класифікації;

φ_{reg} – послідовність операцій гілки регресії.

Такий підхід дає змогу формувати специфічні для кожної підзадачі карти активації, мінімізуючи взаємний негативний вплив градієнтів під час зворотного поширення помилки.

Функціонування гілки класифікації спрямоване на передбачення ймовірнісного розподілу належності виявленого об'єкта до одного з N цільових класів. Вихідний тензор цієї гілки проходить через сигмоїдальну функцію активації, що нормує значення в діапазон [0, 1]. Математично вихідне значення для кожного якоря просторової сітки визначають як умовну ймовірність

$$P = \sigma(W_{cls} \cdot F_{cls} + b_{cls}), \quad (20)$$

де σ – сигмоїдальна функція;

W_{cls} і b_{cls} – навчальні параметри згорткового шару класифікації;

F_{cls} – вхідні ознаки гілки.

В умовах військового застосування, де об'єкти можуть мати візуальну подібність (наприклад різні модифікації танків Т-72 і Т-90), така спеціалізація гілки допомагає мережі краще вивчати ознаки, що відрізняються.

Паралельно функціонує гілка регресії, задачею якої є передбачення геометричних параметрів обмежувальної рамки. Важливою еволюційною зміною в запропонованій моделі є відмова від використання фіксованих якорів (Anchor Boxes) на користь без'якірного підходу (Anchor-free). Замість прогнозування зміщення відносно попередньо заданих шаблонів, модель безпосередньо передбачає відстані від центра поточної чарунки сітки (c_x, c_y) до чотирьох сторін обмежувальної рамки: лівої l, верхньої t, правої r і нижньої b. Вектор виходу регресії

$$(B) = \{l, t, r, b\} \rightarrow \begin{cases} x_{min} = c_x - l \\ y_{min} = c_y - t \\ x_{max} = c_x + r \\ y_{max} = c_y + b \end{cases}, \quad (21)$$

де B – вектор передбачених геометричних параметрів рамки;

l, t, r, b – відстані від центра поточної чарунки сітки до лівої, верхньої, правої та нижньої меж об'єкта відповідно.

Для підвищення точності локалізації в умовах невизначеності (розмиття, часткове перекриття) у гілці регресії застосовано підхід DFL. Замість прогнозування одного детермінованого значення відстані мережа передбачає розподіл імовірностей значень координат. Математичне очікування оцінки координати розраховують інтегральним методом (або дискретною сумою) за розподілом імовірностей $P(y_i)$:

$$\tilde{y} = \sum_{i=0}^{n} P(y_i) \cdot y_i, \quad (22)$$

де $\tilde{y}$ – фінальне оцінене значення координати межі рамки;

$P(y_i)$ – імовірність того, що істинне значення координати знаходиться в точці y_i. І слід розуміти, що $\sum P(y_i) = 1$;

y_i – дискретні значення координат у вікні пошуку.

Така методологія дає змогу враховувати неоднозначність меж об'єкта, що є критичним для обробки відеопотоку з безпілотників, де чіткість контурів часто знижена через вібрації або атмосферні явища. Використання DFL разом із без'якірним підходом значно спрощує архітектуру модуля детекції, зменшує кількість гіперпараметрів і знижує

обчислювальне навантаження на етапі декодування передбачень, роблячи систему більш адаптивною до об'єктів із нестандартним співвідношенням сторін.

Компонент DFL є специфічним для Anchor-free архітектур, які передбачають відстані до сторін рамки. У реальних умовах, особливо зйомка з БПЛА, межі об'єктів часто є розмитими або перекритими, що робить точне визначення координат неоднозначним. DFL допомагає моделювати цю невизначеність, розглядаючи координату не як дискретне число, а як розподіл імовірностей навколо істинного значення. Функція втрат DFL спрямована на те, щоб сконцентрувати (звузити) цей розподіл навколо правильного значення, підвищуючи впевненість мережі в локалізації чітких меж і дозволяючи більшу дисперсію для розмитих меж. Застосування цього підходу значно підвищує стійкість детектора до візуальних спотворень і низької роздільної здатності вхідного зображення.

Процес навчання нейронної мережі базований на мінімізації цільової функції, яка кількісно оцінює розбіжність між передбаченими значеннями та істинною розміткою. У розробленій моделі реалізовано механізм обчислення функції втрат LFn, що являє собою зважену лінійну комбінацію трьох компонентів, кожен із яких відповідає за оптимізацію окремого аспекту детекції. Формально загальна функція втрат описана рівнянням

$$LFn = \lambda_{box} L_{box} + \lambda_{cls} L_{cls} + \lambda_{dfl} L_{dfl}, \quad (23)$$

де L_{box} – втрати регресії координат рамки;

L_{cls} – втрати класифікації об'єкта;

L_{dfl}– фокальні втрати розподілу;

L_{box}, L_{cls}, L_{dfl} – гіперпараметри, що регулюють внесок кожного компонента у загальний градієнт, забезпечуючи баланс між точністю локалізації та розпізнавання.

Для оцінювання якості передбачення геометричних параметрів обмежувальної рамки використовують метрику CIoU (Complete Intersection over Union). На відміну від класичної IoU, яка враховує лише площу перекриття, або DIoU, що додає відстань між центрами, CIoU ураховує три критично важливі геометричні фактори: площу перекриття, евклідову відстань між центрами рамок та узгодженість співвідношення сторін. Математично функція втрат визначена виразом

$$L_{box} = 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{c^2} + \alpha v, \quad (24)$$

де IoU – коефіцієнт перекриття передбаченої (b) та істинної (b^{gt}) рамок;

ρ – евклідова відстань між центрами рамок;

c – довжина діагоналі найменшого прямокутника, що охоплює обидві рамки;

α – ваговий коефіцієнт;

v – параметр, що вимірює подібність відношення сторін $\frac{W}{H}$.

Використання CIoU забезпечує більш стабільну та швидку збіжність форми рамки, запобігаючи осциляціям градієнта на фінальних етапах навчання.

Для навчання гілки класифікації застосовують функцію Varifocal Loss (VFL) або модифіковану бінарну крос-ентропію. Основною проблемою в задачах детекції є екстремальний дисбаланс класів: кількість якорів, що містять фонове зображення, на кілька порядків перевищує кількість якорів з об'єктами. VFL розв'язує цю проблему асиметричним зважуванням прикладів:

$$VFL(p,q) = \begin{cases} -q\left(q\log(p) + (1-q)\log(1-p\right. \\ -\alpha p^{\gamma\log(1-p)}, q = 0 \end{cases} \quad (25)$$

,

де p – передбачена ймовірність класу;

q – цільова оцінка якості для позитивних прикладів (або 0 для негативних);

γ, α – параметри фокусування, що зменшують внесок простих негативних прикладів (фону) у функцію втрат.

Для позитивних прикладів (q > 0) втрата масштабована відповідно до якості локалізації q, що змушує мережу приділяти більше уваги високоякісним детекціям. Для негативних прикладів (q = 0) застосовують фокусуючий параметр, який знижує вагу простих фонових прикладів, запобігаючи домінуванню «легких» негативних зразків у функції втрат.

Для забезпечення практичного використання розробленої нейромережевої моделі в умовах реального часу створено програмну підсистему інференсу (InfE), яка відповідає за розгортання моделі та її взаємодію з джерелами відеосигналу. На етапі експлуатації основними вимогами щодо системи є мінімізація латентності обробки та забезпечення стабільності відеопотоку, чого досягають завдяки модульній об'єктно-орієнтованій архітектурі. Формально підсистема описана кортежем компонентів

$$InfE = \{Grabber, Converter, DeviceMgr, FPS \quad (26)$$

де Grabber – модуль захоплення кадрів;

Converter – блок попередньої обробки зображень;

DeviceMgr – менеджер обчислювальних ресурсів;

FPS_{lim} – механізм стабілізації кадрової частоти.

Центральним елементом підсистеми є клас Grabber, який реалізує патерн проєктування

«Адаптер», створюючи рівень абстракції над низькорівневими інтерфейсами бібліотеки OpenCV. Його основне завдання полягає в інкапсуляції логіки ініціалізації та конфігурації пристроїв захоплення. Клас забезпечує гнучке перемикання між різнорідними джерелами відеосигналу, підтримуючи роботу як із фізичними сенсорами, так і віртуальними потоками, наприклад OBS Virtual Camera. Ця процедура дає змогу динамічно налаштовувати роздільну здатність вхідного потоку для узгодження з параметрами вхідного шару нейромережі, забезпечуючи отримання матриці зображення.

Критично важливим етапом підготовки даних є узгодження колірних моделей, оскільки стандартні драйвери відеопристроїв оперують зображеннями у форматі BGR, а вхідний шар нейронної мережі налаштований на сприйняття простору RGB. Для розв’язання цієї проблеми компонент Converter виконує обов'язкову трансформацію колірного простору вхідного кадру. Математично ця операція описана як перетворення тензора:

$$I_{RGB} = T_{BGR \to RGB}\left(I_{raw}\right), \tag{27}$$

де I_{RGB} – тензор зображення у форматі, необхідному для нейромережі;

I_{raw} – «сире» зображення, отримане з камери (зазвичай BGR);

$\mathcal{T}$ – оператор перестановки каналів масиву.

Паралельно компонент DeviceMgr реалізує механізм автоматичного вибору обчислювального пристрою. Перед завантаженням моделі перевіряють наявність апаратного прискорення:

$$\text{DeviceMgr} = \begin{cases} \text{CUDA, якщо } \exists\ \text{GPU} \\ \text{CPU, інакше} \end{cases}, \tag{28}$$

де CPU (Central Processing Unit) – ідентифікатор режиму виконання обчислень на центральному процесорі;

GPU (Graphics Processing Unit) – логічна умова наявності в системі сумісного графічного прискорювача з підтримкою драйверів CUDA.

У разі виявлення підтримки технології CUDA модель і вхідні тензори автоматично переміщуються у відеопам'ять, що дає змогу використовувати паралелізм графічних ядер для виконання матричних операцій і суттєво пришвидшує процес детекції.

Алгоритм основного циклу обробки побудовано за конвеєрним принципом. На кожній ітерації отриманий кадр передається методу предикції, який виконує пряме проходження нейронної мережі. Результати детекції візуалізують накладанням графічних примітивів. Для забезпечення комфортного сприйняття відеопотоку та синхронізації з частотою оновлення монітора впроваджено компонент FPS_{lim}, який реалізує адаптивний алгоритм стабілізації, що розраховує необхідний час затримки t_{wait} для кожного кадру з метою дотримання цільової частоти FPS_{lim}:

$$t_{wait} = \max\left(0, \frac{1}{FPS_{target}} - \left(t_{end} - t_{start}\right)\right), \tag{29}$$

де t_{wait} – розрахований час затримки перед обробкою наступного кадру;

FPS_{target} – бажана частота кадрів;

t_{end}, t_{start} – системний час завершення та початку циклу обробки поточного кадру відповідно.

Якщо розрахована затримка є додатною, система призупиняє виконання процесу, що дає змогу утримувати стабільну частоту кадрів, запобігаючи розсинхронізації відеоряду та надмірному навантаженню на процесор у моменти простою.

Результати та їх обговорення

Експериментально перевіряли ефективність запропонованої моделі ODS через порівняльний аналіз із базовою архітектурою YOLOv11n на спеціалізованому наборі даних військової техніки. Для забезпечення репрезентативності результатів валідаційну вибірку було розширено до 1318 зображень, що охоплюють 33 класи об'єктів у різних умовах спостереження (день, ніч, складні погодні умови, маскування). Оцінювали якість детекції за метриками mAP50 (середня точність за порогового значення перекриття 50 %) і mAP50-95 (усереднена точність у діапазоні порогів від 50 до 95 %).

На першому етапі досліджень було протестовано базову модель YOLOv11n, яку використовували як точку відліку (baseline). Після 50 епох навчання модель продемонструвала показник mAP50 на рівні 0.603 і mAP50-95 на рівні 0.447. Аналіз похибок показав, що стандартна архітектура ефективно справляється з великими та контрастними об'єктами, проте демонструє зниження чутливості з детекцією малорозмірних цілей та об'єктів зі складним камуфляжем, що підтверджує теоретичне припущення про недостатність локальних рецептивних полів для таких задач.

На другому етапі було досліджено модель із глибокою інтеграцією механізмів уваги (Deep CBAM), де блоки уваги додавали після кожного основного вузла екстрактора ознак. Результати цього експерименту виявилися показовими з точки зору обмежень трансферного навчання. Модель продемонструвала падіння метрики mAP50-95 до рівня 0.246. Такий результат пояснюють явищем «катастрофічного забування» (catastrophic forgetting) і руйнуванням структури попередньо навчених ваг. Масове впровадження нових ініціалізованих шарів призвело до значних змін у розподілі градієнтів, через що 50 епох виявилося недостатньо для відновлення функціональності фільтрів і їхньої адаптації до нових умов. Для коректного навчання такої «важкої»

архітектури необхідна значно більша кількість ітерацій навчання або використання більшого обсягу навчальних даних для стабілізації ваг.

На третьому етапі було реалізовано оптимізовану архітектуру, що отримала умовну назву Lite CA (Lite Coordinate Attention), для якої характерна стратегія селективної інтеграції: замість насичення кожного блока мережі додатковими обчислювальними модулями механізм уваги було впроваджено точково – виключно перед фінальним шаром пірамідальної агрегації. Такий підхід допоміг зберегти семантичну цілісність ознак, сформованих Backbone-частиною, і водночас додати контекстну інформацію. Ця модель продемонструвала найкращий результат: показник mAP50 зріс до 0.651, що на 8 % перевищує базовий рівень. При цьому, незважаючи на незначне зниження метрики строгої локалізації mAP50-95 (0.437), модель показала суттєво вищу здатність до виявлення об'єктів (Recall), що є пріоритетним для військових систем, де пропуск цілі є критичною помилкою.

Для підтвердження коректності процесу навчання та оцінювання динаміки збіжності моделі було проаналізовано зміни основних метрик протягом усіх епох (рис. 1). Графічні залежності демонструють, що функція втрат локалізації (Box Loss) і класифікації стрімко знижується на початкових етапах і виходить на сталий рівень, що свідчить про успішну адаптацію вагових коефіцієнтів мережі. Особливу увагу слід звернути на графік середньої точності (mAP@0.5). Спостерігають стабільне зростання показника без суттєвих коливань, що підтверджує гіпотезу про позитивний вплив інтегрованого механізму уваги на стабільність градієнтів. Високі значення повноти виявлення (Recall), зафіксовані на завершальних етапах, вказують на здатність системи ефективно виявляти об'єкти навіть за умов низької контрастності, мінімізуючи кількість пропусків цілей.

Якісний аналіз ефективності методу наведено на рис. 2, де продемонстровано результат обробки тестового кадру відеопотоку. Як видно із зображення, детектор успішно локалізує об'єкт військової техніки, формуючи обмежувальну рамку з високим ступенем довіри (Confidence Score).

Критично важливим є те, що, незважаючи на наявність природних перешкод і складний фон (рослинність, тіні), межі передбаченої рамки максимально точно відповідають реальним контурам об'єкта. Такої точності позиціонування досягають завдяки роботі модуля Coordinate Attention, який допоміг зберегти просторову інформацію про межі об'єкта після багаторазових операцій зменшення розмірності. Це підтверджує практичну придатність моделі для використання в системах автономного спостереження та розвідки.

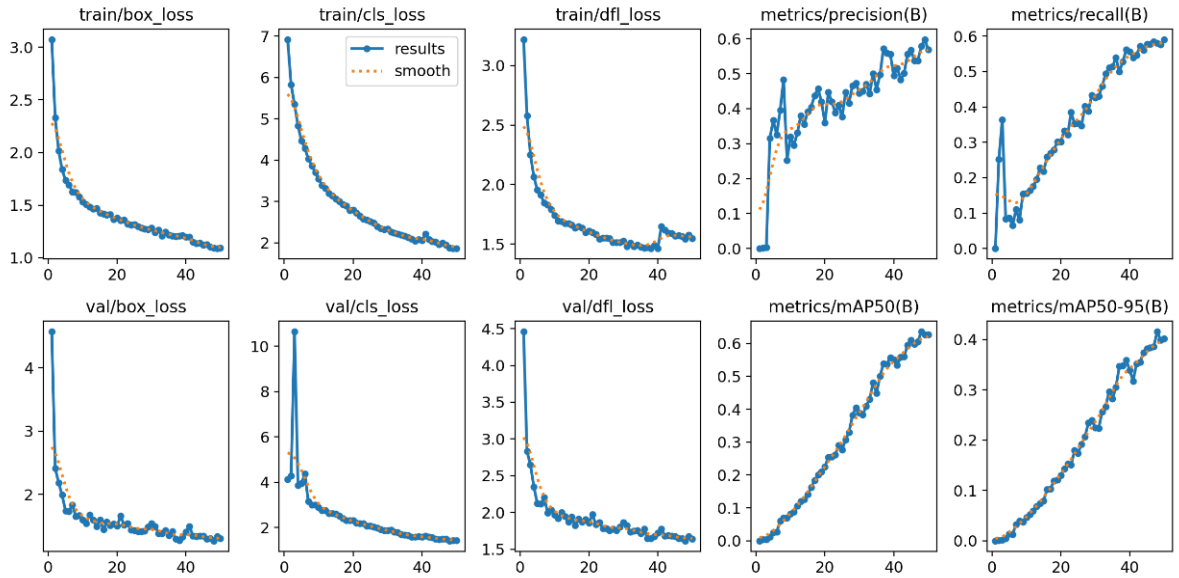

Рис. 1. Динаміка зміни функції втрат і метрик точності (mAP, Recall) під час навчання модифікованої моделі YOLO

На рис. 3 наведено приклад роботи модифікованого детектора на тестовому зображенні, що демонструє результати інференсу в реальних умовах. Як видно з результатів візуалізації, нейромережа успішно локалізувала об'єкт інтересу, сформувавши обмежувальну рамку з високим ступенем довіри (Confidence Score).

Рис. 2. Результати детекції об'єкта бронетехніки в умовах складного фону з відображенням ступеня довіри алгоритму

```
11n-cbam summary (fused): 79 layers, 3,018,958 parameters, 0 gradients, 8.1 GFLOPs
           Class     Images  Instances      Box(P          R      mAP50  mAP50-95): 1
             all       1378       1810      0.633       0.57      0.651      0.437
            T-72A        181        234      0.733      0.637      0.732      0.423
             T-64         52         52       0.71      0.788      0.807      0.535
             T-80         66        112      0.312      0.652      0.407      0.227
             T-90         24         46      0.491       0.63      0.617       0.38
        Leopard 1         24         34      0.485      0.294      0.397      0.234
        Leopard 2        129        203      0.634      0.701      0.742      0.444
             M1A2        105        159      0.605      0.623      0.685      0.435
              XM1         13         13      0.623      0.462      0.596      0.409
        Chalanger         64         80      0.483        0.4      0.441      0.264
             BMP1         31         35       0.64      0.571      0.591      0.425
             BMP2         65         68      0.894      0.809      0.897      0.617
             BMD3         33         33      0.448      0.727      0.656      0.471
  Wolfpack (M1128)         52         52      0.327      0.962       0.92      0.695
            AMX-30         13         13       0.47      0.769      0.752      0.468
           Mercava         33         56      0.584      0.304       0.37       0.23
        Marder 1A2         31         31      0.298      0.452      0.416      0.307
           Leclerc         49         58      0.699      0.586      0.697      0.477
        Bradley M2         82         88      0.735       0.92       0.93       0.67
             Ka-50          7          7          1          0      0.384      0.173
              2S38         68         68      0.781      0.926      0.937      0.671
            Roicat         46         46      0.547      0.196      0.481      0.345
           Type 87         55         55      0.384      0.964      0.843      0.654
               WMA         54         55      0.736      0.636      0.748      0.553
               M60         13         13          1      0.226      0.666      0.453
         Centurion         13         13          1      0.203      0.503      0.376
             UH-1H          9          9          1          0      0.444      0.221
          U-SH 405         44         44      0.517      0.864      0.815      0.612
          Centaora         50         50       0.48        0.7      0.618      0.467
            Vextra         13         13      0.609      0.923      0.885      0.696
             TAM 2         31         31      0.456      0.645      0.597      0.443
            BUC-M3         10         10      0.627        0.9      0.832      0.521
            Tor-m1         10         10      0.581      0.281       0.46      0.267
             AH-1H         19         19          1     0.0537      0.626      0.267
```

Рис. 3. Результат детекції бронетехніки (візуалізація обмежувальної рамки та розрахованого ступеня довіри класифікатора)

Слід відзначити високу щільність прилягання рамки до контурів об'єкта, що свідчить про коректну роботу регресійної частини моделі та мінімізацію похибки локалізації. Незважаючи на наявність візуальних перешкод і низький контраст між об'єктом і фоном (ефект камуфляжу), інтегрований механізм уваги допоміг виділити значущі ознаки цілі, відфільтрувавши інформаційний шум навколишнього середовища. Це підтверджує здатність розробленої архітектури ефективно функціонувати в умовах, наближених до бойових (таблиця).

Таблиця

Порівняння підходів навчання моделей

Архітектура моделі	mAP@0.5	mAP@0.5:0.95	Кількість параметрів (M)	Обчислювальна складність (GFLOPs)	Примітка
YOLOv11n (baseline)	0.603	0.446	2.6	6.3	базова модель без модифікацій
YOLOv11n + Deep CBAM	0.404	0.246	3.5	8.7	важка інтеграція (чотири блоки), перенавчання не вдалося
YOLOv11n + Lite CA	0.651	0.437	3.0	8.1	запропонована модель (один блок CA)

Окремо потрібно зазначити вплив структури датасету на отримані результати. Аналіз покласової точності виявив диспропорцію в результатах: для масових класів (наприклад Bradley M2, 2S38) точність сягала 0.93-0.94, тоді як для рідкісних зразків техніки показники були нижчими. Отримані експериментальні дані свідчать про те, що наявний дисбаланс класів у навчальній вибірці створює додаткові труднощі для навчання, які частково компенсовано впровадженими механізмами уваги, але для подальшого підвищення точності потребують застосування синтетичних даних або розширення датасету.

Висновки

У роботі вирішено науково-практичне завдання підвищення точності автоматизованої детекції військових об'єктів у відеопотоці через удосконалення нейромережевої архітектури.

Аналіз архітектурних рішень показав, що пряма імплементація складних механізмів уваги (важка інтеграція CBAM) у попередньо навчені мережі є неефективною в умовах обмеженого часу навчання (50 епох). Встановлено, що це призводить до дестабілізації ваг моделі та потребує значно більших обчислювальних ресурсів для збіжності, що робить такий підхід недоцільним для швидкого розгортання.

Додатково обґрунтовано, що класичні методи зменшення просторової розмірності (глобальне усереднення) призводять до втрати критично важливої позиційної інформації. Натомість використаний механізм координатної уваги забезпечує декомпозицію процесу просторового кодування, що допомагає зберігати точні координати меж об'єкта навіть із проходженням сигналу через глибокі шари мережі. Це стало головним фактором для покращення локалізації.

Запропонована стратегія точкової інтеграції модуля CA (Lite-версія) допомогла досягти компромісу між складністю моделі та її точністю. Експериментально підтверджено приріст метрики mAP50 до 0.651, що свідчить про підвищення ймовірності виявлення замаскованих і малорозмірних цілей порівняно з базовою архітектурою YOLOv11.

Особливу цінність для прикладної сфери становить зафіксоване зростання повноти виявлення, що свідчить про здатність системи мінімізувати кількість помилок другого роду (пропусків цілі), що є критичним показником для військових задач. Модель продемонструвала високу стійкість до візуальних перешкод, ефективно виділяючи об'єкти на складному фоні в умовах природного маскування.

Також підтверджено ефективність структурної модифікації модуля детекції (Decoupled Head). Розділення потоків обробки для задач класифікації та регресії координат допомогло пришвидшити збіжність алгоритму та усунути конфлікт ознак, що позитивно вплинуло на точність позиціювання обмежувальних рамок.

Розроблена програмна підсистема інференсу забезпечила стабільну роботу моделі в режимі реального часу, реалізуючи адаптивну обробку відеопотоку з різних джерел. Виявлено, що стримуючим фактором для подальшого зростання точності є дисбаланс класів у навчальному наборі даних, що визначає напрям подальших досліджень у площині генерації синтетичних навчальних прикладів. Отримані результати дають змогу рекомендувати розроблену модель для використання в системах ситуаційної обізнаності та автоматизованого цілевказання.

Список використаних джерел

1. Terven J., Cordova-Esparza D., Romero-Gonzalez J. A Comprehensive Review of YOLO Architectures in Computer Vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*. 2023. Vol. 5, Iss. 4. P. 1680-1716. URL: https://doi.org/10.3390/make5040083.

2. Цюник Б. С., Муляревич О. В. Оцінка продуктивності та оптимізація моделей нейронних мереж YOLOv8 для розпізнавання цілей. *Комп'ютерні системи та мережі*. 2024. Т. 6, № 2. С. 242-251. URL: https://doi.org/10.23939/ csn2024.02.242.

3. Назаркевич М., Олексів Н. Система розпізнавання об'єктів на основі моделі YOLO. *Український журнал інформаційних технологій*. 2024. Т. 6, № 1. С. 120-126. URL: https://doi.org/10.23939/ujit2024.01.120.

4. Сліпачук Л., Сліпачук В., Поліщук Т. Побудова системи ідентифікації рухомих об'єктів. *Кібербезпека: освіта, наука, техніка*. 2024. Т. 4, № 24. С. 410-433. URL: https://doi.org/10.28925/2663-4023.2024.25.410433.

5. Слюсар В. І. Застосування нейромережевих технологій для виявлення підводних боєприпасів. *Вісті вищих навчальних закладів. Радіоелектроніка.* 2023. Т. 66, № 3. С. 766-777. URL: https://doi.org/10.20535/S0021347023030020.

6. Zhao Y., Xu S., Wei J. DETRs Beat YOLOs on Real-time Object Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 16965-16974. URL: https://doi.org/10.1109/CVPR52733.2024.01605.

7. Галаган Н. В., Борисенко І. І., Яковець В. П., Бойко О. В. Концептуальна модель системи управління розпізнавання образів із застосуванням ШІ. *Телекомунікаційні та інформаційні технології*. 2025. № 3 (88). URL: https://doi.org/10.31673/2412-4338.2025.038709.

References

1. Terven, J., Cordova-Esparza, D., & Romero-Gonzalez, J. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. *Machine Learning and Knowledge Extraction*, *5*(4), 1680–1716. https://doi.org/10.3390/make5040083

2. Tsiunyk, B. S., & Muliarevych, O. V. (2024). Otsinka produktyvnosti ta optymizatsiia modelei neironnykh merezh YOLOv8 dlia rozpiznavannia tsilei [Performance evaluation and optimization of YOLOv8 neural network models for target recognition]. *Kompiuterni systemy ta merezhi* [*Computer Systems and Networks*], *6*(2), 242–251. https://doi.org/10.23939/csn2024.02.242

3. Nazarkevych, M., & Oleksiv, N. (2024). Systema rozpiznavannia obiektiv na osnovi modeli YOLO [Object recognition system based on the YOLO model]. *Ukrainskyi zhurnal informatsiinykh tekhnolohii* [*Ukrainian Journal of Information Technologies*], *6*(1), 120–126. https://doi.org/10.23939/ujit2024.01.120

4. Slipachuk, L., Slipachuk, V., & Polishchuk, T. (2024). Pobudova systemy identyfikatsii rukhomykh obiektiv [Building a moving object identification system]. *Kiberbezpeka: osvita, nauka, tekhnika* [*Cybersecurity: Education, Science, Technique*], *4*(24), 410–433. https://doi.org/10.28925/2663-4023.2024.25.410433

5. Sliusar, V. I. (2023). Zastosuvannia neiromerezhevykh tekhnolohii dlia vyiavlennia pidvodnykh boieprypasiv [Application of neural network technologies for the detection of underwater ammunition]. *Visti vyshchykh navchalnykh zakladiv. Radioelektronika* [*Radioelectronics and Communications Systems*], *66*(3), 766–777. https://doi.org/10.20535/S0021347023030020

6. Zhao, Y., Xu, S., & Wei, J. (2024). DETRs beat YOLOs on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16965–16974). https://doi.org/10.1109/CVPR52733.2024.01605

7. Halahan, N. V., Borysenko, I. I., Yakovets, V. P., & Boiko, O. V. (2025). Kontseptualna model systemy upravlinnia rozpiznavannia obraziv iz zastosuvanniam ShI [Conceptual model of a pattern recognition control system using AI]. *Telekomunikatsiini ta informatsiini tekhnolohii* [*Telecommunication and Information Technologies*], (3(88)). https://doi.org/10.31673/2412-4338.2025.038709

T. Filimonchuk, Y. Koltun, M. Maslov

Adaptive Model for Equipment Recognition with Attention Mechanisms Based on YOLOv11

Abstract. Background. *The relevance of the study is driven by the need to improve the efficiency of computer vision systems in modern warfare, where the rapid development of unmanned aerial vehicles and reconnaissance equipment requires reliable tools for the automated detection of military hardware in real time. Despite their high processing speed, existing detectors demonstrate a decrease in accuracy when dealing with small, camouflaged objects or in complex background environments due to the loss of spatial information during convolutional operations. Therefore, improving the neural network architecture model by integrating attention mechanisms and optimizing computational processes is of particular importance.* ***Object of the study*** *is the process of automated detection and localization of military objects in an optical video stream.* ***Subject of the study*** *encompasses methods for the structural modification of the YOLO convolutional neural network architecture using coordinate attention mechanisms and decoupled detection modules.* ***Results.*** *The paper proposes and implements a modified neural network architecture model, the key feature of which is the pinpoint integration of a lightweight coordinate attention module (Lite CA) before the feature pyramid aggregation layer. This allowed for the decomposition of the spatial encoding process, preserving precise positional information that is typically lost during standard global averaging. The introduction of the «detection head»*

component into the model ensured the independent processing of classification and coordinate regression tasks. It was experimentally confirmed that the adaptive model achieves a mean average precision (mAP@0.5) of 0.651, which is 8 % higher than the baseline, and demonstrates a significant increase in recall. ***Conclusions.*** *The proposed approach provides an effective balance between localization accuracy and computational complexity, guaranteeing the detector's high robustness against visual interference and camouflage effects. The developed model is recommended for implementation in autonomous situational awareness and target designation systems, allowing for the minimization of target omission errors in combat conditions.*

Keywords: *YOLOv11, CBAM, attention mechanism, object detection, computer vision, transfer learning, convolutional neural networks, Python, PyTorch.*

Філімончук Тетяна Володимирівна, *кандидат технічних наук, доцент кафедри «Електронних обчислювальних машин», Харківський національний університет радіоелектроніки, Харків, Україна, E-mail: tetiana.filimonchuk@nure.ua, ORCID: 0000-0002-4380-504X*

Колтун Юрій Миколайович, *кандидат технічних наук, доцент кафедри «Інформаційно-мережної інженерії» Харківський національний університет радіоелектроніки, Харків, Україна, E-mail: yurii.koltun@nure.ua, ORCID: 0000-0003-2680-9978*

Маслов Микола Костянтинович, магістрант, Харківський національний університет радіоелектроніки, Харків, Україна, *E-*mail: mykola.maslov@nure.ua, *ORCID: 0009-0000-1653-3408*

Tetiana Filimonchuk, *PhD, Associate Professor of the Department of Electronic Computers, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, E-mail: tetiana.filimonchuk@nure.ua, ORCID: 0000-0002-4380-504X*

Yuriy Koltun*, PhD, Associate Professor of the Department of Information and Network Engineering, Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, E-mail, e-mail: yurii.koltun@nure.ua, ORCID: 0000-0003-2680-9978*

Maslov Mykola, master's student, *Kharkiv National University of Radio Electronics, Kharkiv, Ukraine, E-mail:* mykola.maslov@nure.ua, *ORCID: 0009-0000-1653-3408*

Стаття надійшла 11.03.26
Стаття прийнята до друку після рецензування 20.04.26
Стаття опублікована (оприлюднена) 29.05.26