https://doi.org/10.31891/2219-9365-2026-86-41
УДК 004.932.72:004.032.26

ГОЗАК Ярослав
Київський національний університет імені Тараса Шевченка
https://orcid.org/0009-0008-2963-7204
e-mail: hozakya@fit.knu.ua

ROI-SSD: ОДНОЕТАПНИЙ ДЕТЕКТОР ІЗ ПІДТРИМКОЮ РЕГІОНІВ ІНТЕРЕСУ

У цій роботі запропоновано ROI-SSD — розширення архітектури Single Shot Detector (SSD), орієнтоване на обробку регіонів інтересу та призначене для детекції об'єктів на зображеннях довільного розміру. Запропонований метод забезпечує обробку локальних фрагментів зображення замість повного кадру за рахунок використання динамічної генерації базових обмежувальних рамок та адаптивного скорочення глибини згорткової частини мережі залежно від масштабу регіону інтересу.

На відміну від традиційних детекторів, що навчаються та оцінюються на зображеннях фіксованого розміру, запропонований підхід явно враховує невідповідність між навчальними даними та умовами застосування моделі при роботі з частковими фрагментами сцени. Для усунення цієї проблеми запропоновано стратегію прогресивного багаторозмірного навчання, яка передбачає поступове розширення множини допустимих роздільностей вхідних даних та включення кропів навколо об'єктів.

Експериментальні результати показують, що запропонована модифікація архітектури зберігає точність детекції на повних зображеннях порівняно з базовою моделлю SSD. Водночас моделі, навчені лише на повних зображеннях, демонструють суттєве зниження точності при застосуванні до часткових фрагментів сцени. Запропонована стратегія навчання дозволяє підвищити точність детекції на регіонах інтересу у кілька разів, що підтверджує важливість узгодження умов навчання з цільовим сценарієм використання.

Отримані результати свідчать про те, що ROI-SSD є практичним розширенням SSD для детекції на регіонах довільного розміру та може слугувати основою для реалізації ROI-орієнтованої обробки у системах з обмеженими обчислювальними ресурсами.

Ключові слова: нейронні мережі розпізнавання зображень, регіони інтересу.

HOZAK Yaroslav
Taras Shevchenko National University of Kyiv

ROI-SSD: REGION OF INTEREST SINGLE SHOT DETECTOR

This paper proposes ROI-SSD, an extension of the Single Shot Detector (SSD) architecture designed to support object detection on image regions of arbitrary size and aspect ratio. Unlike conventional approaches that process full frames of fixed resolution, the proposed method enables detection on local image fragments. This is achieved through dynamic generation of default bounding boxes and adaptive truncation of the convolutional backbone depending on the scale of the corresponding region of interest.

Conventional object detectors are typically trained and evaluated on fixed-size images, which leads to a significant degradation in detection accuracy when applied to partial image regions. In this work, it is shown that the primary cause of this degradation is the mismatch between the training data distribution and the actual inference conditions. To address this issue, a progressive multi-resolution training strategy is introduced. This strategy gradually expands the set of input resolutions during training and incorporates cropped regions around objects at later stages.

Experimental results demonstrate that the proposed architectural modification preserves detection accuracy on full images compared to the baseline SSD model. At the same time, models trained only on full images exhibit a substantial drop in performance when applied to cropped regions. The proposed training strategy significantly improves detection accuracy on partial image regions, confirming the importance of aligning training conditions with the target inference scenario.

The results indicate that ROI-SSD provides a practical extension of SSD for detection on arbitrarily sized regions and forms a foundation for ROI-aware processing in applications with constrained computational resources.

Keywords: neural networks for object detection, regions of interest.

Стаття надійшла до редакції / Received 27.03.2026
Прийнята до друку / Accepted 28.04.2026
Опубліковано / Published 31.05.2026

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Системи автоматичного виявлення об'єктів на зображеннях та відео є важливим компонентом сучасних інформаційних систем комп'ютерного зору. Такі системи широко застосовуються в задачах відеоспостереження, автономної навігації, робототехніки, аналізу відеопотоків та обробки даних дистанційного зондування. Значного прогресу в цій галузі було досягнуто завдяки використанню згорткових нейронних мереж, серед яких особливе місце займають одностадійні детектори об'єктів, такі як SSD (Single Shot Detector)[1], YOLO[2] та їх модифікації.

Класичні архітектури детекції об'єктів, зокрема SSD, зазвичай передбачають використання вхідних зображень фіксованої роздільної здатності. У більшості реалізацій це досягається шляхом масштабування зображення до квадратної форми з подальшою зміною співвідношення сторін або використанням доповнення

(letterbox). Такий підхід спрощує реалізацію мережі, але водночас обмежує можливості застосування моделі у задачах, де обробка здійснюється лише на частині сцени або на регіонах інтересу (Region of Interest, ROI).

У багатьох практичних системах комп'ютерного зору обробка повного кадру є надлишковою з обчислювальної точки зору, оскільки об'єкти зазвичай займають лише невелику частину зображення. Тому перспективним підходом є регіонально-орієнтована обробка, при якій нейронна мережа застосовується лише до вибраних фрагментів зображення. Однак використання стандартних детекторів у такому режимі є проблематичним, оскільки більшість з них розраховані на фіксований розмір входу та не призначені для обробки зображень довільної форми та роздільності.

Одним із можливих шляхів вирішення цієї проблеми є модифікація архітектури детектора таким чином, щоб забезпечити можливість обробки вхідних зображень довільного розміру. У цій роботі розглядається модифікована архітектура SSD, яка дозволяє виконувати детекцію об'єктів на зображеннях та фрагментах довільної роздільності без необхідності попереднього приведення їх до фіксованого формату. Така модифікація створює передумови для побудови регіонально-орієнтованих систем детекції, у яких обробка здійснюється лише на тих ділянках сцени, де очікується наявність об'єктів.

Разом з тим, ефективність детекції на часткових фрагментах сцени значною мірою залежить від розподілу даних, використаного під час навчання моделі. Якщо мережа навчається виключно на повних зображеннях, її здатність до коректної роботи на частково видимих сценах або невеликих регіонах може суттєво знижуватись. Тому важливим завданням є розробка методів навчання, які підвищують стійкість детектора до змін масштабу, співвідношення сторін та обмеженого просторового контексту.

У цій роботі запропоновано підхід до навчання моделі, що базується на прогресивному розширенні множини допустимих роздільностей вхідних зображень. На ранніх етапах навчання модель тренується на зображеннях стандартного розміру, після чого поступово вводяться приклади з різними розмірами та співвідношеннями сторін, включаючи часткові кропи сцени. Така стратегія дозволяє підвищити стійкість моделі до обробки регіонів інтересу та зменшити деградацію точності при зменшенні доступного просторового контексту.

Метою роботи є дослідження можливості побудови фреймворку регіонально-орієнтованої детекції на основі модифікованої архітектури SSD та аналіз впливу різних стратегій масштабування і формування регіонів на точність детекції об'єктів. Експериментальні результати демонструють, що використання прогресивного багаторозмірного навчання дозволяє суттєво підвищити точність детекції на часткових фрагментах сцени та забезпечити стабільну роботу моделі при значному зменшенні площі оброблюваного регіону.

АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Дослідження ефективних методів детекції об'єктів у відео охоплюють декілька напрямів, які певною мірою пов'язані з цілями цієї роботи, проте принципово відрізняються підходами до використання регіонів інтересу (ROI), повторного використання ознак та організації динамічних обчислень.

Детекція об'єктів у відео та часово-просторова агрегація ознак

Класичні методи детекції об'єктів у відео виконують обробку повних кадрів і переважно зосереджені на часовій агрегації ознак або результатів детекції. Наприклад, метод Flow-Guided Feature Aggregation (FGFA) виконує деформацію та агрегацію ознак із сусідніх кадрів за допомогою оптичного потоку, що дозволяє підвищити точність детекції, однак при цьому згорткова мережа обробляє кожен повний кадр [3]. У Spatio-Temporal Sampling Networks (STSN) використовуються навчувані зміщення вибірки у просторі та часі для отримання релевантних ознак із сусідніх кадрів [4]. Метод MEGA застосовує модуль глобально-локальної агрегації з пам'яттю, що інтегрує інформацію з довготривалого часового контексту [5].

Інший напрямок представлений підходами на основі tubelet-послідовностей. Так, Tubelet Proposal Networks (TPN) генерують короткі просторово-часові траєкторії об'єктів («tubelets»), які надалі класифікуються згортковими мережами [6]. Метод SELSA виконує агрегацію семантичної інформації на рівні об'єктних треків для уточнення результатів детекції [7], тоді як DRNet використовує динамічні рекурентні мережі для поширення інформації між кадрами [8]. Усі ці підходи передбачають формування повної піраміди ознак для кожного кадру та використання часової надлишковості відео. Водночас вони не виконують просторового розрідження обчислень і не обмежують обробку лише регіонами інтересу, визначеними за попередніми кадрами.

Динамічні та розріджені обчислення

Інший напрям досліджень пов'язаний із використанням динамічних обчислювальних графів та умовних обчислень у згорткових нейронних мережах. Наприклад, у SkipNet мережа навчається пропускати окремі залишкові блоки залежно від вхідних даних, що дозволяє зменшити кількість операцій без значної втрати точності [9]. У [10] запропонували підхід, у якому під час інференсу адаптивно вимикається частина фільтрів згорткових шарів. У [11] досліджують адаптивні нейронні мережі, що вибирають підмоделі різної складності залежно від складності вхідних даних.

Метод Sparse R-CNN пропонує іншу стратегію зменшення обчислювальних витрат: замість щільної

сітки якорів використовується невелика кількість навчуваних пропозицій, які уточнюються детектором у повністю наскрізному режимі [12]. Хоча ці роботи дозволяють скоротити кількість активних параметрів або обчислювальних гілок, вони все одно виконують обробку повних карт ознак для всього зображення. Таким чином, вони не розбивають зображення на окремі фрагменти та не адаптують просторову область обробки або глибину мережі залежно від розміру регіону інтересу чи інформації з попередніх кадрів.

Детекція з фокусом на регіонах та дизайн якорів

У низці сучасних детекторів використовуються більш складні механізми уточнення регіонів. Наприклад, RefineDet застосовує модуль уточнення якорів, після якого виконується детекція об'єктів, що дозволяє поступово покращувати локалізацію об'єктів [13]. У FoveaBox відмовляються від традиційної системи якорів і передбачають межі об'єктів на основі так званих фовеальних областей, зосереджених у центральних частинах об'єктів [14].

Попри використання більш складних механізмів роботи з регіонами та якорями, такі методи все одно працюють із глобальними картами ознак і не розглядають регіони інтересу як окремі фрагменти зображення із власними картами ознак та динамічно сформованими наборами якорів. Зокрема, вони не підтримують режим, у якому через мережу пропускаються лише фрагменти зображення, що перекриваються з відстежуваними об'єктами.

Ефективні архітектури для ресурсно-обмежених систем

Окремий напрям досліджень присвячений розробці архітектур детекторів, оптимізованих для мобільних та вбудованих систем. Наприклад, EfficientDet використовує масштабування архітектури та ефективні модулі об'єднання ознак для зменшення обчислювальних витрат [15]. Детектори на основі MobileNetV3 застосовують легкі згорткові блоки та оптимізовані голови детекції для досягнення роботи в реальному часі на мобільних пристроях [16]. Однак у таких моделях зменшення обчислювальної складності є переважно статичним: однаковий обсяг обчислень виконується для кожного кадру та для кожної області зображення незалежно від кількості або розміру об'єктів.

Позиціонування запропонованого підходу

На відміну від розглянутих методів, у цій роботі запропоновано регіонально-орієнтоване розширення архітектури SSD для обробки відео (ROI-SSD), яке передбачає:

1. використання результатів детекції на попередніх кадрах для формування невеликої множини прямокутних регіонів інтересу;
2. можливість об'єднання близько розташованих об'єктів у більші регіони на основі моделі вартості, що враховує компроміс між кількістю проходів мережі та сумарною площею регіонів;
3. виконання згорткової частини мережі лише для цих регіонів, причому глибина обробки може змінюватися залежно від масштабу об'єкта;
4. динамічну генерацію наборів базових обмежувальних рамок (default boxes) для кожного регіону залежно від просторової роздільності його карти ознак;
5. проєкцію результатів детекції, отриманих на рівні окремих регіонів, у глобальну систему координат кадру з подальшою часовою агрегацією.

Наскільки відомо авторам, у наявних роботах з детекції об'єктів у відео не поєднуються одночасно такі властивості: (i) просторове розрідження обчислень шляхом обробки лише окремих фрагментів зображення; (ii) адаптивне виконання згорткової мережі залежно від масштабу регіону; (iii) динамічна генерація базових обмежувальних рамок для кожного регіону в архітектурі, подібній до SSD. Це дозволяє розглядати запропонований підхід як новий напрям побудови систем детекції об'єктів, орієнтованих на обробку регіонів інтересу та ефективну роботу у ресурсно-обмежених середовищах. Загальні характеристики моделей та їх особливості наведені у таблиці 1.

Як видно з таблиці, більшість сучасних методів детекції у відео зосереджені на часовій агрегації ознак або на зменшенні обчислювальної складності мережі, проте вони виконують обробку повного кадру. На відміну від них, запропонований підхід поєднує просторове розрідження обчислень із використанням регіонів інтересу, адаптивну глибину згорткової мережі та динамічну генерацію якорів для кожного регіону.

Останні дослідження також розглядають адаптивні стратегії обробки, спрямовані на зосередження обчислень на найбільш інформативних ділянках зображення або на динамічну адаптацію компонентів нейронної мережі. Наприклад, метод AdaFocus обирає регіони інтересу для обробки з підвищеною роздільною здатністю у задачах розпізнавання відео [17]. Підхід Dynamic Head вводить адаптивні механізми уваги, що дозволяють змінювати конфігурацію голови детекції залежно від контексту ознак [18]. Детектори на основі трансформерів, такі як DETR, формулюють задачу детекції об'єктів як задачу передбачення множини об'єктів і повністю відмовляються від використання якорів [19].

Таблиця 1

Ключові характеристики та особливості існуючих моделей розпізнавання об'єктів

Метод	**Домен**	**Повний кадр чи ROI/тайли**	**Динамічні обчислення (глибина / фільтри)**	**Робота з якорями / пропозиціями**	**Відношення до запропонованого підходу**
FGFA	Відео	Повний кадр	Ні	Стандартні якорі (Faster R-CNN / SSD)	Повторне використання ознак у часі; не використовує просторове розрідження або ROI
STSN	Відео	Повний кадр	Ні	Стандартна щільна детекція	Просторово-часова агрегація, але обробка повного кадру
MEGA	Відео	Повний кадр	Ні	Стандартна щільна детекція	Використовує довготривалий часовий контекст, але без ROI-орієнтованої обробки
TPN / Tubelet CNN	Відео	Повний кадр	Ні	Tubelet-пропозиції	Використовує tubelets, але обчислення виконуються для повних кадрів
SELSA	Відео	Повний кадр	Ні	Стандартні пропозиції регіонів	Семантична агрегація ROI у часі, але без тайлової обробки
DRNet	Відео	Повний кадр	Обмежено (рекурентна мережа)	Стандартна щільна детекція	Часова оптимізація, але без умовного виконання мережі для ROI
SkipNet	Зображення	Повний кадр	Так (пропуск residual-блоків)	Будь-яка схема	Динамічна глибина, але однакова для всього зображення
Skip Filters	Зображення	Повний кадр	Так (пропуск фільтрів)	Будь-яка схема	Динамічна ширина мережі, без просторового розділення на ROI
Adaptive NN	Зображення	Повний кадр	Так (вибір підмережі)	Будь-яка схема	Адаптивна складність, але обробка повного зображення
Sparse R-CNN	Зображення	Повний кадр	Ні	Розріджені навчувані пропозиції	Розрідженість у просторі пропозицій, але не у просторі зображення
RefineDet	Зображення	Повний кадр	Ні	Двоетапне уточнення якорів	Уточнення регіонів, але без тайлової обробки
FoveaBox	Зображення	Повний кадр	Ні	Без якірне	Орієнтація на центр об'єкта, але щільна обробка всього зображення
EfficientDet	Зображення / відео	Повний кадр	Ні (статичне масштабування)	Якірне	Статична оптимізація обчислень без ROI
MobileNetV3	Зображення / відео	Повний кадр	Ні (статична архітектура)	Переважно якірне	Ефективні для edge-пристроїв, але завжди обробляють повний кадр
ROI- SSD (запропонований підхід)	Відео	**ROI / патч** (прямокутні частини зображення навколо об'єктів)	**Так (глибина мережі залежить від масштабу ROI)**	Динамічна генерація якорів	Поє днує часове повторне використання ROI, просторове розрідження обчислень та адаптивну глибину мережі

Запропонований метод

Ми пропонуємо ROI-SSD — нове розширення архітектури SSD для детекції об'єктів у відео, яке використовує часову стабільність положення об'єктів та виконує обробку лише для невеликої кількості локальних регіонів інтересу (Region of Interest, ROI). Кожен такий регіон обробляється згортковою частиною мережі з адаптивно скороченою глибиною, після чого до отриманих карт ознак застосовується модифікована SSD-голова з динамічно сформованими наборами базових рамок (default boxes), специфічними для кожного ROI. Це дозволяє реалізувати просторово розріджену та масштабно-адаптивну обробку, що суттєво зменшує обчислювальні витрати при збереженні точності детекції.

Загальний конвеєр обробки запропонованого підходу проілюстровано на рис. 1.

Рис. 1. Запропонована структура динамічного SSD з урахуванням регіонів інтересу. Виявлення з попереднього кадру визначають області інтересу, які обробляються незалежно з адаптивною глибиною згорткової мекрежі та генерацією специфічних для регіону якорів.

Запропонований підхід складається з чотирьох основних компонентів:

1. пропозиція регіонів базуючись на детекціях з попередніх кадрів,
2. об'єднання близько розташованих об'єктів у спільні регіони;
3. адаптивне виконання згорткової частини мережі залежно від масштабу ROI;
4. динамічна генерація наборів базових рамок (default boxes) для кожного ROI.

Загальний алгоритм роботи складається з послідовної обробки кадрів, де для кожного нового кадру виконуються такі кроки:

1. Отримуються результати детекції на попередньому кадрі B_i;
2. Кожна обмежувальна рамка розширюється до відповідного регіону інтересу R_i
3. Регіони інтересу кластеризуються та об'єднуються відповідно до критерію, що базується на моделі вартості.
4. Для кожного об'єднаного регіону інтересу виконуються такі операції:

a. регіон передається для обробки через перші L_Retапів згорткової частини мережі;

b. для відповідних карт ознак генеруються базові обмежувальні рамки (default boxes), специфічні для даного регіону інтересу;

c. SSD-голови детекції застосовуються до кожної карти ознак;

d. отримані передбачення декодуються та перетворюються до глобальної системи координат кадру.

5. Результати детекції для всіх регіонів інтересу об'єднуються, після чого за потреби застосовується часовий фільтр або алгоритм відстеження об'єктів.

Запропонований конвеєр використовує просторове розрідження обчислень і масштабно-адаптивну обробку, які відсутні у класичній архітектурі SSD. Завдяки цьому обчислювальні витрати можуть суттєво зменшуватися у випадках, коли на сцені присутня лише невелика кількість об'єктів.

Формування регіонів інтересу

Нехай детекції на попередньому кадрі t−1 представлені множиною обмежувальних рамок

$$B_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i}), i = 1, \dots, N, \tag{1}$$

де x_1, y_1, x_2, y_2— координати обмежувальної рамки відповідного об'єкта.

Для кожної обмежувальної рамки формується контекстно-розширений регіон інтересу шляхом її розширення.

$$\begin{aligned} p_{xi} &= \alpha_w (x_{2i} - x_{1i}) + \delta_x, \\ p_{yi} &= \alpha_h (y_{2i} - y_{1i}) + \delta_y, \end{aligned} \tag{2}$$

де α_w, α_hє відносними коефіцієнтами доповнення та δ_x, δ_yмінімальні відступи на рівні пікселів.

Розширений прямокутний регіон інтересу:

$$R_i = (\max(0, x_{1i} - p_{xi}), \max(0, y_{1i} - p_{yi}), \min(W, x_{2i} + p_{xi}), \min(H, y_{2i} + p_{yi})). \tag{3}$$

Регіон R_i задається у вигляді прямокутника, вирівняного за осями координат, що узгоджується з вимогами операторів згорткових нейронних мереж

Об'єднання регіонів інтересу

У випадках, коли декілька об'єктів розташовані близько один до одного, незалежна обробка кожного ROI може призводити до надлишкових обчислень. Наприклад, два сусідні об'єкти можуть формувати два майже однакові регіони інтересу, що потребують окремих проходів мережі.

Для зменшення кількості обчислень у роботі використовується процедура об'єднання ROI. Якщо два регіони R_i та R_j перекриваються або знаходяться на невеликій відстані один від одного, вони можуть бути замінені одним більшим регіоном:

$$R_{ij} = \mathrm{u}nion(R_i, R_i) \tag{4}$$

Рішення про об'єднання приймається на основі простої моделі вартості, яка враховує компроміс між кількістю проходів мережі та сумарною площею оброблюваних регіонів. Якщо об'єднання двох ROI зменшує загальні обчислювальні витрати, такі регіони об'єднуються.

Таким чином, на кожному кадрі формується обмежена множина регіонів

$$R_t = \{R_1, R_2, \dots, R_k\} \tag{5}$$

Нехай

$$A_i = \text{area}(R_i) \tag{6}$$

є площею R_i. Для кожної пари регіонів інтересу (R_i, R_j), об'єднаний регіон інтересу визначається як найменший прямокутник, що вміщує в себе R_i, R_j:

$$R_{ij} = \text{bbox}(R_i \cup R_j), \tag{7}$$

із площею

$$A_{ij} = \text{area}(R_{ij}). \tag{8}$$

Припустимо, що обчислювальна вартість виконання скороченої згорткової частини мережі для регіону інтересу R_i :

$$\text{cost}(R) = c_{\text{pix}} \cdot A_R + K, \tag{9}$$

де

c_{pix} – ціна за піксель (FLOPs масштабовані за глибиною),

K– це постійні накладні витрати на запуск згорткової мережі та голови SSD для одного регіону інтересу.

Критерій злиття:

Злиття R_i та R_j є виграшним коли:

$$\text{cost}(R_{ij}) < \text{cost}(R_i) + \text{cost}(R_j), \tag{10}$$

що дає:

$$c_{\text{pix}} A_{ij} + K < c_{\text{pix}}(A_i + A_j) + 2K. \tag{11}$$

Можна переформулювати як:

$$(A_i + A_j - A_{ij}) > \tau, \tau = \frac{K}{c_{\text{pix}}}. \tag{12}$$

Таким чином, об'єднання регіонів є доцільним лише у випадку, якщо «зекономлена площа» перевищує порогове значення τ. Оскільки A_{ij} визначається як площа мінімального прямокутника, що охоплює обидва регіони, така умова природним чином накладає штраф на випадки, коли регіони інтересу просторово рознесені в обох вимірах.

Для реалізації цього підходу використовується жадібний алгоритм агломеративного кластерування, який послідовно об'єднує пари регіонів інтересу з додатним виграшем площі доти, доки таких пар більше не залишається

Динамічна глибина згорткової мережі для кожного регіону інтересу

Традиційна SSD обробляє повне зображення та виконує всі шари згорткової мережі та всі рівні піраміди ознак. На відміну від цього, наш метод обробляє кожен регіон інтересу незалежно та скорочує згорткову мережу залежно від масштабу ROI.

Нехай регіон інтересу має висоту h_Rта ширину w_R. Визначимо характеристичний розмір:

$$s_R = \sqrt{h_R \cdot w_R}. \tag{13}$$

Нехай етапи згорткової частини мережі SSD позначено індексами $\{1, \dots, L\}$, Кожному етапу відповідає крок дискретизації $\{S_1, \dots, S_L\}$ та карта ознак розміру (H_l, W_l).

Маленькі регіони інтересу потребують лише ранніх шарів; більші регіони вимагають глибших шарів. Визначимо відображення:

$$L_R = \max\{l: S_l \le s_R/\gamma\}, \tag{14}$$

де γ є гіперпараметром (емпірично $\gamma \in [2,4]$) що визначає, скільки пікселів має відображатися принаймні в одній комірці ознаки.

Регіон інтересу передається для обробки лише через шари $1, \dots, L_R$. У цьому випадку обчислювальна вартість визначається як:

$$\text{cost}(R) \approx \sum_{l=1}^{L_R} k_l \, H_l(R) \, W_l(R), \tag{15}$$

де k_l — коефіцієнт обчислювальної вартості згорткового шару, а $H_l(R)$, $W_l(R)$ змінюються пропорційно до розмірів регіону інтересу.

Така схема забезпечує масштабно-адаптивне скорочення глибини згорткової частини мережі, що недоступно для детекторів, які виконують обробку повного кадру.

Динамічне створення рамок за замовчуванням для довільних карт ознак регіонів інтересу

На відміну від стандартного SSD, наші регіони інтересу мають довільну роздільну здатність та створюють карти ознак довільних розмірів.$(H_l(R), W_l(R))$. Оригінальна сітка якорів SSD не може бути використана.

Таким чином, для кожного регіону інтересу та для кожного збереженого рівня ознак l, динамічно формуються базові обмежувальні рамки (default boxes) відповідно до такого правила:

$$\begin{aligned} c_x &= (j + 0.5)/W_l(R), \\ c_y &= (i + 0.5)/H_l(R), \end{aligned} \tag{16}$$

для $i = 0 \dots H_l(R) - 1$, $j = 0 \dots W_l(R) - 1$, разом із параметрами масштабу, аналогічними до тих, що використовуються в SSD:

$$w = s_k\sqrt{a_r}, h = s_k/\sqrt{a_r}, \tag{17}$$

і шкала середнього геометричного:

$$s'_k = \sqrt{s_k s_{k+1}}. \tag{18}$$

Оскільки голови SSD є згортковими, прогнози працюють правильно для карти ознак будь-якого розміру.

Після отримання прогнозів у системі координат регіону інтересу, рамки масштабуються та зміщуються назад до глобальних координат зображення.:

$$\begin{aligned} x_{\text{global}} &= x_{\text{roi}} + x_{1R}, \\ y_{\text{global}} &= y_{\text{roi}} + y_{1R}. \end{aligned} \tag{19}$$

Стратегія навчання

Ефективність детекції об’єктів на часткових фрагментах сцени значною мірою залежить від розподілу даних, використаного під час навчання моделі. Більшість сучасних детекторів об’єктів навчаються на повних зображеннях фіксованої роздільної здатності, що призводить до зниження точності при застосуванні моделі до фрагментів зображення або регіонів інтересу довільної форми. У випадку регіонально-орієнтованої обробки така невідповідність між навчальним і тестовим розподілом може призводити до суттєвої деградації точності детекції.

З метою підвищення стійкості моделі до обробки регіонів інтересу в роботі використовується стратегія прогресивного багаторозмірного навчання, яка поєднує навчання на повних зображеннях стандартної роздільної здатності з поступовим введенням прикладів різних масштабів і співвідношень сторін.

Початкова фаза навчання

На початковому етапі навчання модель тренується на повних зображеннях, приведених до стандартного розміру 300×300. Масштабування виконується із застосуванням схеми letterbox, яка зберігає співвідношення сторін зображення та доповнює його до квадратної форми шляхом додавання відступів.

Цей етап дозволяє стабілізувати процес навчання та забезпечує формування базових ознак для детекції об’єктів різних класів. Використання повних зображень на ранніх епохах також забезпечує достатній просторовий контекст для формування високорівневих ознак.

Прогресивне розширення множини роздільностей

Після початкової фази навчання застосовується підхід, у якому множина допустимих роздільностей вхідних зображень поступово розширюється. На кожній наступній групі епох до навчальної вибірки

додаються нові розміри вхідних зображень.

Нехай S(e) — множина допустимих роздільностей на епосі e. На початку навчання $S(1) = \{300 \times 300\}$,

, а з ростом номера епохи множина розширюється: $S(e) \subseteq S(e+1)$,

До множини роздільностей включаються як квадратні, так і прямокутні формати, наприклад: 256×256,144×144,32×256,256×32

Для кожного навчального прикладу роздільність вхідного зображення випадково вибирається з множини $S(e)$. Такий підхід дозволяє моделі поступово адаптуватися до обробки зображень різних масштабів та співвідношень сторін.

Важливою особливістю цієї схеми є те, що початкові роздільності не виключаються з навчання, а лише доповнюються новими. Таким чином, модель продовжує отримувати приклади повних зображень навіть на пізніх етапах навчання, що запобігає деградації точності при обробці повного кадру.

Навчання на підзображеннях регіонів інтересу

На пізніх етапах навчання до навчальної вибірки додаються приклади, отримані шляхом формування кропів навколо об'єктів. Для кожної обмежувальної рамки об'єкта формується регіон інтересу шляхом розширення рамки на певний відступ.

Такий кроп може містити як один, так і декілька об'єктів. Метою використання подібних прикладів є навчання моделі працювати в умовах обмеженого просторового контексту, що відповідає режиму роботи запропонованого фреймворку.

При формуванні кропів накладається обмеження на мінімальний відступ між межею об'єкта та межею зображення. Нехай d — мінімальна відстань між рамкою об'єкта та межею кропу. У роботі використовується умова $d \geq 8$ пікселів. Це обмеження запобігає утворенню вироджених прикладів, у яких об'єкт розташований безпосередньо на межі зображення.

Переваги запропонованої стратегії навчання

Запропонована стратегія навчання дозволяє значно підвищити стійкість моделі до змін масштабу та форми вхідних даних. Поєднання навчання на повних зображеннях та часткових фрагментах сцени забезпечує узгодження між розподілом навчальних і тестових даних.

Зокрема, використання прогресивного багаторозмірного навчання дозволяє:

- підвищити точність детекції на регіонах інтересу малого розміру;
- зменшити деградацію точності при зменшенні просторового контексту;
- зберегти високу точність детекції на повних зображеннях.

Експериментальні результати, наведені у наступному розділі, демонструють, що запропонована стратегія навчання суттєво покращує точність детекції на часткових фрагментах сцени порівняно з моделями, навченими лише на повних зображеннях.

Експериментальні дослідження

Експериментальні дослідження

Метою експериментальних досліджень є оцінювання ефективності запропонованого підходу ROI-SSD та аналіз впливу різних стратегій масштабування і формування регіонів інтересу на точність розпізнавання об'єктів. Особлива увага приділяється дослідженню здатності моделі працювати з фрагментами зображення різного розміру, що є ключовою вимогою для реалізації регіонально-орієнтованої обробки відео.

Експерименти включають порівняння класичної архітектури SSD та запропонованого підходу ROI-SSD у різних умовах тестування, а також дослідження впливу різних стратегій навчання на точність розпізнавання.

Налаштування експерименту

Усі моделі навчалися та оцінювалися на підмножині testval датасету детекції об'єктів VOC. Якість детекції оцінювалася за метрикою mean Average Precision (mAP@0.5).

Було проведено кілька груп експериментів:

1. порівняння SSD та ROI-SSD при обробці повних зображень;
2. аналіз впливу різних методів масштабування зображення;
3. дослідження точності детекції на часткових фрагментах сцени;
4. оцінювання впливу запропонованої стратегії прогресивного багаторозмірного навчання.

Порівняння SSD та ROI-SSD

Спочатку було виконано порівняння класичної моделі SSD та модифікованої моделі ROI-SSD у стандартному режимі обробки повних зображень.

Результати показали, що запропонована модифікація архітектури практично не впливає на точність детекції.

Таблиця 2

Порівняння SSD та ROI-SSD під час обробки повних кадрів. Запропонована архітектура зберігає точність виявлення, водночас забезпечуючи висновок на основі ROI

Модель	Масштабування під час тренування	Масштабування під час валідації	mAP
SSD	класичне	класичне	0.797
SSD	класичне	letterbox	0.781
ROI-SSD	класичне	класичне	0.734
ROI-SSD	класичне	letterbox	0.721

Результати показали, що запропонована модифікація архітектури не призводить до суттєвої деградації точності при обробці повних зображень. Значення mAP для ROI-SSD є близькими до відповідних значень для базової моделі SSD, що підтверджує коректність узагальнення архітектури на випадок вхідних даних довільної роздільності.

Цей результат є важливим, оскільки демонструє, що підтримка обробки регіонів інтересу не вимагає компромісу у точності при роботі з повними зображеннями.

Вплив методу масштабування

У наступній групі експериментів досліджувався вплив різних методів масштабування вхідних зображень. Зокрема, порівнювалися два підходи:

- класичне масштабування із зміною співвідношення сторін;
- масштабування із збереженням співвідношення сторін із доповненням (letterbox).

Результати показали, що зміна методу масштабування може призводити до певного зниження точності. Наприклад, для SSD модель, навчена на класичному масштабуванні, показала mAP=0.7814 при тестуванні на letterbox-зображеннях.

Аналогічна тенденція спостерігається і для ROI-SSD. Це свідчить про те, що узгодженість політики масштабування між етапами навчання та тестування є важливим фактором для досягнення максимальної точності.

Детекція на часткових фрагментах сцени

Ключовою метою запропонованого підходу є можливість виконання детекції на часткових фрагментах зображення. Тому було проведено серію експериментів, у яких модель ROI-SSD застосовувалася до кропів навколо об’єктів із різними значеннями відступу (padding).

Експерименти показали, що модель, навчена лише на повних зображеннях, демонструє різке зниження точності при застосуванні до кропів навколо об’єктів. Зокрема, при зменшенні відступу до 20 пікселів значення mAP знижується до ~0.08, що свідчить про сильну залежність моделі від глобального контексту сцени.

Це підтверджує, що стандартне навчання на повних зображеннях не забезпечує узагальнення на задачі детекції в обмеженому полі зору.

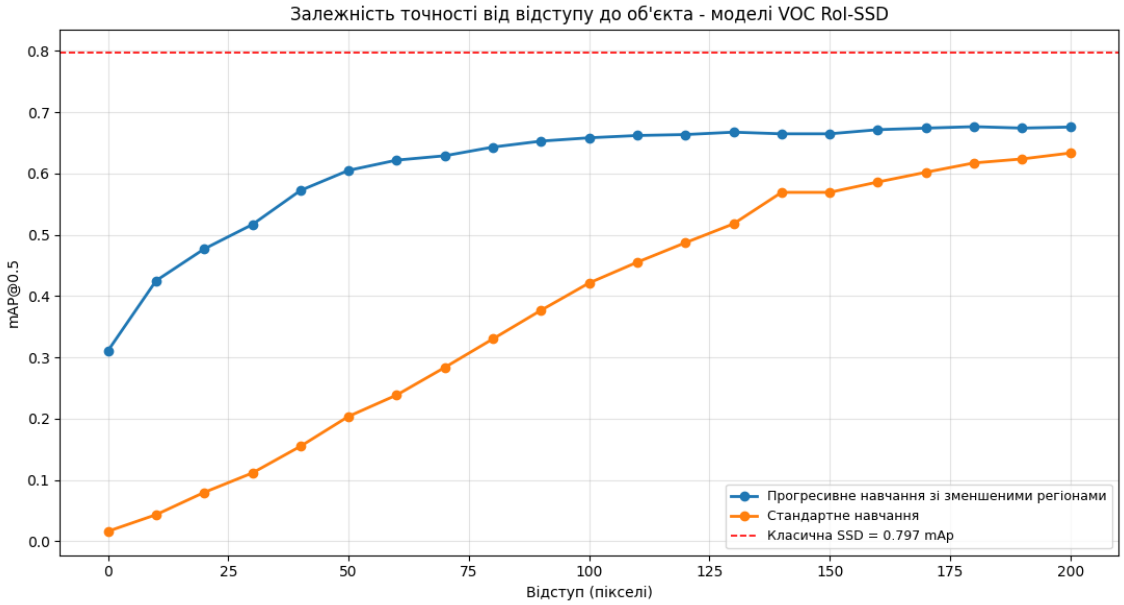

Рис. 2. Точність виявлення як функція доповнення регіонів інтересу

Як показано на рис. 2, базова модель, навчена лише на повних зображеннях, демонструє різке падіння точності, коли просторовий контекст стає обмеженим.

Ці результати свідчать про суттєву залежність точності детекції від доступного просторового контексту та розподілу навчальних даних.

Вплив прогресивного багаторозмірного навчання

Для підвищення стійкості моделі до обробки часткових фрагментів сцени було використано запропоновану стратегію прогресивного багаторозмірного навчання, описану в попередньому розділі.

Таблиця 3

Вплив прогресивного навчання з кількома роздільними здатностями на точність виявлення для обрізаних областей

Стратегія тренування	Відступ 20 пікселів	Відступ 50 пікселів	Відступ 100 пікселів	Повний кадр (зменшений до 300 по довшій стороні зі збереженням пропорцій)	Повний кадр (300x300)
Класичне тренування	0.08	0.20	0.42	0.66	**0.786**
Прогресивне навчання зі зменшеними регіонами	**0.480**	**0.60**	**0.66**	**0.67**	0.734

У таблиці 2 підсумовано покращення точності виявлення, отримане за допомогою запропонованої стратегії навчання для різних значень доповнення регіонів інтересу.

У цьому випадку модель на пізніх етапах навчання отримувала приклади з різними розмірами зображень і співвідношеннями сторін, включаючи кропи навколо об’єктів. Використання такої стратегії призвело до значного покращення точності детекції на часткових фрагментах сцени.

Запропонована стратегія прогресивного багаторозмірного навчання дозволяє суттєво підвищити точність детекції на часткових фрагментах сцени. Зокрема, при padding 20 пікселів точність зростає з ~0.08 до ~0.51 mAP, що відповідає більш ніж шестикратному покращенню.

Аналогічне покращення спостерігається і для більших значень padding, що свідчить про зменшення залежності моделі від розміру доступного контексту.

Отримані результати підтверджують, що ключовим фактором ефективності детекції на ROI є узгодження навчального розподілу з умовами використання моделі.

Аналіз результатів

Результати експериментів показують, що запропонована архітектурна модифікація SSD забезпечує коректну роботу детектора для зображень і регіонів довільного розміру без втрати точності на повному кадрі. Водночас основним обмеженням при переході до ROI-обробки є невідповідність між навчальним і тестовим розподілами.

Запропонована стратегія навчання ефективно усуває цю проблему, дозволяючи моделі адаптуватися до умов обмеженого просторового контексту. У результаті детектор стає значно більш стійким до змін масштабу та форми вхідних даних..

Обговорення результатів

Отримані результати дозволяють зробити кілька висновків щодо властивостей запропонованої архітектури ROI-SSD та впливу стратегії навчання на її здатність працювати з регіонами інтересу довільного розміру. Насамперед встановлено, що модифікація базової архітектури SSD, яка включає динамічну генерацію базових рамок і адаптивну глибину згорткової частини мережі, не призводить до суттєвої деградації точності при обробці повних зображень. Це свідчить про те, що запропонована архітектура зберігає основні властивості базового детектора та може розглядатися як його узагальнення на випадок вхідних даних довільної роздільності.

Разом із тим результати експериментів показують, що сама по собі архітектурна модифікація не гарантує високої точності при переході до часткових фрагментів сцени. Модель, навчена лише на повних зображеннях, демонструє різке зниження mAP при застосуванні до кропів навколо об’єктів, особливо за малих значень відступів. Це означає, що основним обмеженням у режимі ROI-обробки є не стільки архітектура детектора, скільки невідповідність між навчальним розподілом і умовами використання моделі.

Запропонована стратегія прогресивного багаторозмірного навчання істотно зменшує цю невідповідність. Поступове розширення множини допустимих роздільностей, поєднане з включенням кропів навколо об’єктів, дозволяє моделі адаптуватися до змін масштабу, співвідношення сторін та обмеженого просторового контексту. Експериментальні результати показують, що використання такої стратегії приводить до багаторазового підвищення точності на часткових фрагментах сцени порівняно з базовим навчанням лише на повних зображеннях. Отже, саме узгодження режиму навчання з цільовим режимом використання є ключовим фактором для ефективної роботи ROI-SSD.

Окремо слід відзначити вплив просторового контексту. Експерименти з різними значеннями відступу при формуванні кропів показали, що зменшення видимої частини сцени знижує точність детекції, однак після прогресивного багаторозмірного навчання ця залежність стає значно слабшою. Це свідчить про те, що модель навчається ефективніше використовувати локальні ознаки об’єкта та менше залежати від глобального

контексту повного кадру.

Таким чином, отримані результати підтверджують доцільність поєднання архітектурної модифікації SSD із спеціалізованою стратегією навчання. Запропонований підхід забезпечує коректну роботу детектора на зображеннях і регіонах довільного розміру, а також істотно підвищує стійкість до часткового спостереження сцени. Це дозволяє розглядати ROI-SSD як придатну основу для побудови регіонально-орієнтованих систем детекції, у яких обсяг обчислень може адаптуватися до просторової структури вхідних даних.

Водночас слід зазначити, що в межах цієї роботи основна увага приділялася архітектурі моделі та стратегії її навчання. Тому питання кількісної оцінки обчислювальної ефективності, а також дослідження повного відеоконвеєра з міжкадровим формуванням і супроводженням регіонів інтересу потребують окремого подальшого аналізу.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У роботі запропоновано ROI-SSD — модифікацію архітектури SSD, що забезпечує можливість детекції об’єктів на зображеннях і регіонах довільного розміру. Основними компонентами підходу є динамічна генерація базових рамок та адаптивне скорочення глибини згорткової частини мережі залежно від масштабу регіону інтересу.

Експериментальні результати показали, що запропонована архітектура зберігає точність детекції на повних зображеннях на рівні базової моделі SSD. Водночас встановлено, що стандартне навчання на повних зображеннях не забезпечує ефективної роботи на часткових фрагментах сцени.

Для вирішення цієї проблеми запропоновано стратегію прогресивного багаторозмірного навчання, яка дозволяє узгодити навчальний і тестовий розподіли. Використання цієї стратегії забезпечує суттєве підвищення точності детекції на регіонах інтересу, особливо в умовах обмеженого просторового контексту.

Отримані результати демонструють, що ефективна робота детекторів на регіонах довільного розміру потребує не лише архітектурних змін, але й відповідної організації процесу навчання. Запропонований підхід створює основу для подальших досліджень у напрямку ROI-орієнтованої обробки зображень і відео, зокрема з урахуванням обчислювальної ефективності та міжкадрової узгодженості..

References

1. Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C.-Y., Berg A. C. SSD: Single shot multibox detector // Computer Vision – ECCV 2016 : proceedings of the 14th European Conference on Computer Vision. Cham : Springer, 2016. P. 21–37. DOI: 10.1007/978-3-319-46448-0_2.

2. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016. P. 779–788. DOI: 10.1109/CVPR.2016.91.

3. Zhu X., Wang Y., Dai J., Yuan L., Wei Y. Flow-guided feature aggregation for video object detection // Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, 2017. P. 408–417. DOI: 10.1109/ICCV.2017.51.

4. Bertasius G., Torresani L., Shi J. Object detection in video with spatiotemporal sampling networks // Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018. P. 331–346. DOI: 10.1007/978-3-030-01216-8_20.

5. Chen Y., Cao Y., Hu H., Wang L., Lin S., Dai J. MEGA: Memory enhanced global-local aggregation for video object detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020. P. 10307–10316. DOI: 10.1109/CVPR42600.2020.01032.

6. Kang K., Ouyang W., Li H., Wang X. Object detection from video tubelets with convolutional neural networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, 2017. P. 817–825. DOI: 10.1109/CVPR.2017.93.

7. Wu H., Chen Y., Wang N., Zhang Z. Sequence level semantics aggregation for video object detection // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019. P. 9217–9225. DOI: 10.1109/ICCV.2019.00931.

8. Li X., Wang W., Hu X., Yang J. Learning temporal information for video object detection via dynamic recurrent neural networks // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019. P. 3530–3539. DOI: 10.1109/CVPR.2019.00365.

9. Wang X., Yu F., Dou Z.-Y., Darrell T., Gonzalez J. E. SkipNet: Learning dynamic routing in convolutional networks // Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018. P. 409–424. DOI: 10.1007/978-3-030-01216-8_25.

10. Chen T., Goodfellow I., Shlens J. NetAdapt: Platform-aware neural network adaptation for mobile applications // Proceedings of the European Conference on Computer Vision (ECCV). Munich, 2018. P. 285–300. DOI: 10.1007/978-3-030-01216-8_18.

11. Bolukbasi T., Wang J., Dekel O., Saligrama V. Adaptive neural networks for efficient inference // Proceedings of the International Conference on Machine Learning (ICML). Sydney, 2017. P. 527–536.

12. Sun P., Zhang R., Jiang Y., Kong T., Xu C., Zhan W., Tomizuka M., Li L., Yuan Z., Wang C., Luo P. Sparse R-CNN: End-to-end object detection with learnable proposals // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021. P. 14454–14463. DOI: 10.1109/CVPR46437.2021.01422.

13. Zhang S., Wen L., Bian X., Lei Z., Li S. Z. Single-shot refinement neural network for object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, 2018. P. 4203–4212. DOI: 10.1109/CVPR.2018.00442.

14. Kong T., Sun F., Liu H., Jiang Y., Li L., Shi J. FoveaBox: Beyond anchor-based object detection // IEEE Transactions on Image Processing. 2020. Vol. 29. P. 7389–7398. DOI: 10.1109/TIP.2020.2992036.

15. Tan M., Pang R., Le Q. V. EfficientDet: Scalable and efficient object detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020. P. 10781–10790. DOI: 10.1109/CVPR42600.2020.01079.

16. Howard A., Sandler M., Chu G., Chen L.-C., Chen B., Tan M., Wang W., Zhu Y., Pang R., Vasudevan V., Le Q. V., Adam H. Searching for MobileNetV3 // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019. P. 1314–1324. DOI: 10.1109/ICCV.2019.00140.

17. Wu Z., Xiong Y., Yu S. X., Lin D. AdaFocus: Towards end-to-end adaptive video recognition // Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020. P. 171–180. DOI: 10.1109/CVPR42600.2020.00025.

18. Dai X., Chen Y., Xiao B., Chen D., Liu M., Yuan L. Dynamic head: Unifying object detection heads with attentions // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021. P. 7373–7382. DOI: 10.1109/CVPR46437.2021.00729.

19. Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. End-to-end object detection with transformers // Computer Vision – ECCV 2020. Cham : Springer, 2020. P. 213–229. DOI: 10.1007/978-3-030-58452-8_13.

20. Hozak Y., Paliy S. Modern small networks for image classification. Feature analysis // Управління розвитком складних систем. Київ : КНУБА, 2024. № 60. С. 221–229. DOI: 10.32347/2412-9933.2024.60.221-229.

21. Гозак Я., Палій С. Порівняльний аналіз нейронних мереж розпізнавання об'єктів на зображеннях // Вісник Хмельницького національного університету. Технічні науки. 2025. Т. 357, № 5.1. С. 86–98. DOI: 10.31891/2307-5732-2025-357

22. Hozak Y., Paliy S. Optimizing Kernel Configurations and Convolutional Strategies for Efficient Shallow CNNs in Real-Time Vision Systems. SIST 2025. IEEE 5th International Conference on Smart Information Systems and Technologies Proceedings, 2025. DOI: 10.1109/SIST61657.2025.11139312