Автоматизовані системи управління та прилади автоматики. 2025. № 4 (187)

УДК 004.8 **DOI:** https://doi.org/10.30837/0135-1710.2025.187.005

Г. Влах-Вигриновська, В. Кромкач

ВИБІР МЕТОДИКИ ОЦІНЮВАННЯ ТОЧНОСТІ ДЛЯ ЗАВДАНЬ АНАЛІЗУ СТАТИЧНИХ СЦЕН НА ОСНОВІ ЗГОРТКОВИХ НЕЙРОННИХ МЕРЕЖ

Об’єктом вивчення є методики кількісного оцінювання точності й надійності прогнозів згорткових нейронних мереж у завданнях аналізу статичних сцен, зокрема семантична сегментація та монокулярне метричне оцінювання глибини. Використані теоретичні, аналітичні та емпіричні наукові **методи дослідження**: порівняльний аналіз, синтез, систематизація, експериментальне моделювання тощо. **Актуальність роботи** зумовлена тим, що традиційні метрики оцінювання точності не завжди беруть до уваги особливості завдань аналізу статичних сцен – дисбаланс класів, локалізацію та малі об’єкти, шум чи змінне освітлення. Це знижує точність результатів і потребує впровадження гібридних метрик, граничноорієнтованих і метрик кількісного оцінювання невизначеності (UQ) для забезпечення надійності й безпеки систем. **Метою дослідження** є обґрунтування й вибір найбільш ефективної та доцільної методики оцінювання точності для завдань аналізу статичних сцен на основі згорткових нейронних мереж способом порівняння й систематизації наявних метрик, аналізу їх переваг і обмежень у різних класах завдань і розроблення інтегрованого фреймворку для підвищення якості оцінювання. Для досягнення окресленої мети необхідно виконати такі завдання: провести порівняльний аналіз традиційних метрик; дослідити сучасні підходи й вибір релевантних метрик і протоколів для конкретних класів завдань; розробити концептуальний гібридний фреймворк, що забезпечує повну валідацію моделі, зважаючи на перекриття, геометричну точність і калібрування впевненості. Унаслідок дослідження сформульовано висновки. Для аналізу статичних сцен оптимально комбінувати такі метрики: *accuracy* й *F*1 – з метою класифікації, *IoU* і *mAP* – для детекції. Найбільш ефективні – *mAP* для складних сцен і для детекції малих об’єктів. Запропоновано гібридний фреймворк, що забезпечує повну валідацію моделі, покриваючи загальний об’єм, геометричну якість і надійність. Цей фреймворк поєднує гранично-орієнтовані метрики для забезпечення геометричної точності та методологію кількісного оцінювання невизначеності для калібрування впевненості та локалізації помилок. Це розв’язує проблему невідповідності між високою точністю моделей та обмеженістю стандартних метрик валідації. Перехід до *Boundary IoU* та метрик відстані, зокрема *Hausdorff Distance*, забезпечить масштабно-збалансовану та значно вищу чутливість до помилок на контурах, слугуватиме інструментом для виявлення катастрофічних локальних геометричних відхилень. Концептуальний фреймворк стимулює розроблення більш надійних і точних архітектур ЗНМ.

Ключові слова: комп’ютерний зір; семантична сегментація; детекція об’єктів; згорткові нейромережі; метрики оцінювання.

1. Вступ

Швидкий розвиток архітектур згорткових нейронних мереж (ЗНМ) і їх широке застосування в завданнях аналізу зображень (семантична сегментація, детекція, класифікація сцен) призвів до ситуації, коли результати різних робіт важко порівнювати через розбіжності в наборах даних, метриках і протоколах оцінювання. Автори сучасних досліджень 2022–2025 рр. наголошують на необхідності стандартизації підходів до

оцінювання, інтеграції метрик для калібрування й оцінювання невизначеності, а також у *cross-dataset*-бенчмарках, що моделюють реальні доменні зсуви.

Точність оцінювання відіграє важливу роль, оскільки статичні сцени часто містять складні елементи, такі як малі об'єкти, шум або варіативне освітлення. Проблема полягає в тому, що традиційні метрики оцінювання, наприклад загальна точність, не завжди беруть до уваги особливості завдань, як-от дисбаланс класів або локалізацію об'єктів.

Згідно з останніми тенденціями перехід від класичних метрик до спеціалізованих, зокрема *mAP* та *IoU*, стає необхідним для об'єктивного порівняльного оцінювання моделей ЗНМ (*CNN*). Це особливо актуально в умовах швидкого розвитку глибокого навчання, де нові архітектури вимагають адаптованих методів оцінювання для забезпечення високої продуктивності й узагальнення.

Особливо високі вимоги до точності висуває монокулярне метричне оцінювання глибини (*MMDE*). Тоді як традиційні методи часто прогнозують лише відносну глибину, *MMDE* прагне до отримання карт глибини з абсолютною шкалою. Цей перехід є необхідним для забезпечення геометричної узгодженості, що дає змогу надійно розгортати системи в реальному світі без додаткового калібрування. З огляду на критичні застосування, де висока точність і надійність є життєво важливими (наприклад, у самокерованих транспортних засобах або медичній діагностиці), методики оцінювання продуктивності ЗНМ мають бути не менш досконалими, ніж самі моделі [1].

2. Аналіз літературних джерел і постановка проблеми дослідження

Останні дослідження щодо оцінювання *CNN* для аналізу статичних сцен зосереджені на вдосконаленні метрик, зважаючи на особливості завдань, зокрема на детекцію малих об'єктів у складних сценах чи класифікацію зображень. Наприклад, у роботі [2], присвяченій детекції малих цілей у складних статичних сценах, автори пропонують мережу *IIHNet*, оцінюючи її за допомогою прецизії, повноти, *mAP* і *AP*50, демонструючи покращення на 20–70 %, якщо порівнювати з базовими моделями, такими як *Faster RCNN*.

Аналогічно в дослідженні впливу архітектурних факторів *CNN* на класифікацію зображень віддаленого зондування впроваджено метрики загальної точності (*OA*), кількості параметрів та *FLOPs*, показано, що збільшення глибини мережі покращує семантичне навчання, але надмірна глибина знижує *OA* [3].

Серед загальних метрик для *CNN* в завданнях аналізу зображень виокремлюють *accuracy, precision, recall, F1-score, mAP* і *IoU* як основні для класифікації, детекції та сегментації.

Літературний огляд точності в *CNN* для віддаленого зондування виявив перехід від традиційних метрик (*OA, Kappa*) до *DL*-специфічних, таких як *F1-score* (використовується в 59 % досліджень для бінарних завдань) та *mAP* (29 %), з акцентом на *P-R*-криві для завдань з порогами. У дослідженні класифікації сцен на датасеті *Places2* застосовується *top*-5 *accuracy* як основна метрика через неоднозначність міток, з *ResNet*-34 досягаючи 38.57 % *top*-1 *accuracy* [4–6]. Дослідники візуальних методів пояснюваності *CNN* упроваджують нові метрики, зокрема *Average Drop* та *Insertion / Deletion scores,*

для оцінювання правильності пояснень, пов'язаних із точністю моделі в завданнях класифікації зображень. Унаслідок вивчення останніх наукових праць можемо виокремити детальні дослідження калібрування та невизначеності, розвідки, що аналізують помилки метрик і їх інтерпретацію в реальних завданнях, серед яких огляди з калібрування [7], великі систематичні дослідження невизначеності та *UQ* [8], якості наборів даних і нових *ReM*-версій *COCO* [9], а також статті, присвячені помилковим інтерпретаціям метрик у зображеннях. Загалом тенденції 2022–2025 рр. вказують на інтеграцію метрик пояснюваності з традиційними для кращого розуміння поведінки мереж.

У сфері семантичної сегментації постійний розвиток енкодер-декодер архітектур спрямований на підвищення точності, особливо на стиках об'єктів. Дослідники розробляють методи, які посилюють внутрішню узгодженість об'єктів і забезпечують "загострення границь". Існують також підходи, що поєднують глибокі ЗНМ з умовно випадковими полями (*CRF*) для ефективного захоплення контекстуальних зв'язків між семантичними ознаками й ознаками глибини, що демонструє прагнення до покращення точності на рівні моделювання [10].

У контексті оцінювання глибини прогрес від відносної до абсолютної метричної глибини (*MMDE*) є вирішальним для забезпечення геометричної узгодженості, необхідної для 3D-реконструкції та навігації. Оскільки ці завдання вимагають високоточного просторового оцінювання, методи валідації мають забезпечувати виявлення будь-яких геометричних неточностей, що можуть призвести до збоїв у реальних системах [11].

Основний недолік полягає в *boundary un-awareness* (нечутливість до границь). Цей ефект зумовлений тим, що різниця в точному позиціюванні або формі границі несуттєво впливає на загальний бал, доки площа перекриття залишається незмінною. Традиційні метрики не здатні адекватно оцінювати моделі, оптимізовані для роботи зі складними структурами й дрібними гранулярними елементами.

Недосконаліть традиційних метрик перекриття для об'єктивного оцінювання геометричної точності ЗНМ, поряд із зростанням потреби в кількісному оцінюванні надійності прогнозу, формує ключову проблему дослідження. Необхідність удосконалення методик оцінювання точності аналізу статичних сцен на основі ЗНМ визначається двома основними напрямами, які потрібно виконати одночасно, а саме:

- розроблення або адаптація метрик, які забезпечують чутливість до граничних помилок і топології, водночас замінити або доповнити обмежені метрики перекриття;
- інтеграція методології кількісного оцінювання невизначеності (*UQ*) для оцінювання надійності та калібрування впевненості.

3. Мета й завдання дослідження

Метою цієї роботи є обґрунтування та вибір найбільш ефективної та доцільної методики оцінювання точності для завдань аналізу статичних сцен на основі згорткових нейронних мереж способом порівняння та систематизації наявних метрик, аналізу їх переваг і обмежень у різних класах завдань і розроблення інтегрованого фреймворку для підвищення якості оцінювання.

Для досягнення окресленої мети необхідно розв’язати такі завдання:

- провести порівняльний аналіз переваг і обмежень ключових метрик, таких як *accuracy, precision, recall, F1-score, IoU* та *mAP*, у контексті різних архітектур *CNN*;

- дослідити сучасні підходи й обрати релевантні метрики й протоколи для конкретних класів завдань;

- розробити концептуальний гібридний фреймворк, що забезпечує повну валідацію моделі з огляду на перекриття, геометричну точність і калібрування впевненості.

4. Виклад основного матеріалу

Аналіз статичних сцен за допомогою *CNN* передбачає завдання класифікації, сегментації та детекції об’єктів, де оцінка точності є критичною. Основні методики оцінювання можна поділити на метрики, орієнтовані на класифікацію (наприклад, *accuracy, precision, recall, F1-score*), і метрики, орієнтовані на локалізацію (наприклад, *Intersection over Union* (*IoU*) та *mean Average Precision* (*mAP*)) з адаптаціями для конкретних архітектур *CNN*.

Рис. 1. Метрики оцінювання точності

Традиційні метрики достатні для збалансованих наборів даних у завданнях класифікації, проте незбалансовані або просторово складні сцени в статичному аналізі вимагають гібридних протоколів, що містять як глобальні, так і попіксельні оцінки [12]. *Accuracy* вимірює частку правильних прогнозів серед усіх зразків, забезпечуючи простий глобальний індикатор продуктивності. Її перевага полягає в простоті та інтерпретації для збалансованих завдань класифікації з використанням архітектур, таких як *ResNet*, де вона ефективно оцінює загальну надійність моделі в маркуванні статичних сцен (наприклад, міських середовищ). Однак *accuracy* дає збій у незбалансованих наборах даних, типових

для аналізу сцен, де домінантні класи (наприклад, фонові пікселі) завищують показники, маскуючи невдачі на рідкісних об'єктах (як приклад, пішоходи).

Precision (TP/(TP + FP)) і *Recall* (TP/(TP + FN)) розв'язують окреслені проблеми, зосереджуючись на продуктивності позитивного класу. Зокрема *precision* перевершує в мінімізації помилкових тривог у завданнях виявлення з *YOLO* або *Faster R-CNN*, тоді як *recall* забезпечує всебічне охоплення елементів сцени. Обмеження передбачають чутливість до порогів упевненості та незбалансованості класів, що може призводити до компромісів у реальному часі статичного виявлення [13].

F1-score – це гармонійне середнє між *precision* та *recall*, що балансує їх, використовується в 40–59 % досліджень для мультикласових завдань. Метрика *F1-score* пропонує стійкість для незбалансованих статичних сцен у класифікаторах або детекторах *CNN* і є особливо корисною для оцінювання варіантів *U-Net* у семантичній сегментації, де коефіцієнт *Dice* (еквівалент *F1* для бінарних випадків) кількісно оцінює перетин, досягаючи високих показників у наборах даних міського керування. Однак *F1-score* ігнорує просторові нюанси, обмежуючи його корисність в архітектурах, що вимагають точності меж, таких як *CNN* на базі розширених конволюцій (наприклад, *DeepLab*).

IoU оцінює просторовий перетин (перетин / об'єднання) між передбаченими й реальними регіонами, що є критичним для сегментації в статичних сценах. Обмеження передбачають нечутливість до помилок меж і помилкових позитивів поза регіонами перетину, роблячи *IoU* менш ідеальним для завдань з пріоритетом на виявлення.

mAP агрегує середню точність по класах та порогах *IoU*, перевершуючи у виявленні об'єктів з *YOLOv8*, де фіксує варіації масштабів у статичних сценах (наприклад, *mAP* 0.88 у порівняльних дослідженнях). Недоліком *mAP* є обчислювальна інтенсивність і залежність від вибору порогу, що може недооцінювати моделі в розріджених сценах.

У табл. 1 систематизовано інформацію про метрики, які найчастіше використовуються, їх переваги й недоліки.

Таблиця 1. *Порівняльна характеристика традиційних метрик*

Метрика	**Застосування**	**Переваги**	**Недоліки**	**Рекомендоване застосування в архітектурах CNN**
Accuracy	Класифікація	Проста, інтуїтивна; добре для збалансованих даних	Вразлива до незбалансованості; маскує помилки в рідкісних класах	*ResNet* для класифікації сцен (наприклад, *ADE20K*)
Precision	Детекція	Мінімізує помилкові позитивні; корисна для уникнення помилкових тривог	Ігнорує помилкові негативні; залежить від порогу	*YOLO/Faster R-CNN* для виявлення в спостереженні

Кінець таблиці 1

Метрика	**Застосування**	**Переваги**	**Недоліки**	**Рекомендоване застосування в архітектурах CNN**
Recall	Детекція	Забезпечує виявлення всіх об'єктів; критична для повноти	Може призводити до багатьох помилкових позитивних	*U-Net* для сегментації в медичному аналізі
F1-score	Сегментація	Балансує *precision* та *recall*; стійка до незбалансованості	Ігнорує просторові деталі; припускає рівну важливість	Гібридні моделі для імбалансованих наборів
IoU	Локалізація	Вимірює просторову точність; стійка до імбалансу класів	Нечутлива до помилок меж; не для класифікації	*HRNet/PSPNet* для семантичної сегментації
mAP	Комплексні сцени	Агрегує по класах та порогах; всебічна для багатокласових завдань	Обчислювально складна; залежить від вибору *IoU*	*YOLOv8* для виявлення об'єктів у статичних сценах

Вибір метрик також залежить від архітектури *CNN*, яка використовується для аналізу статичних сцен. Наприклад, для завдань класифікації зображень, де потрібно визначити, до якого класу належить зображення, часто використовують *accuracy, precision, recall* та *F1-score*.

Для завдань сегментації зображень, де потрібно розділити зображення на різні ділянки, застосовують *IoU* та інші метрики, основані на піксельній точності.

Для завдань детекції об'єктів, де потрібно виявити та локалізувати об'єкти на зображенні, використовується *mAP*.

Для класифікації в статичних сценах (наприклад, маркування сцен з *ResNet*) найбільш релевантні *accuracy* та *F1-score*, доповнені *precision/recall* для імбалансу; протоколи передбачають крос-валідацію на наборах даних, наприклад *ADE20K*. Завдання виявлення об'єктів (наприклад, *YOLO* на наборах типу *COCO* для статичних зображень) пріоритизують *mAP* з порогами *IoU*, даючи змогу аналізувати помилки через криві *precision-recall* для критичних застосувань, як-от спостереження.

Семантична сегментація (наприклад, *U-Net* на *CamVid*) рекомендує *mIoU* та *Dice/F1* з протоколами, що містять точність переднього плану для класів, орієнтованих на безпеку, та час інференсу для реального часу.

Гібридні протоколи, які поєднують *mAP/IoU* для паноптичних завдань, з'являються для всебічного розуміння сцен, забезпечуючи узгодження метрик з обмеженнями розгортання, наприклад *FPS* в автономних системах [14, 15].

У табл. 2 подано інформацію щодо використання метрик і протоколів оцінювання для різного класу завдань.

Таблиця 2. *Релевантні метрики та протоколи для конкретних класів завдань*

Клас завдання	**Рекомендовані метрики**	**Протоколи оцінювання**	**Приклади архітектур *CNN***
Класифікація зображень	*Accuracy, F1-score, Precision/Recall*	Крос-валідація, аналіз імбалансу на *ADE20K*	*ResNet, VGGNet*
Виявлення об’єктів	*mAP, IoU, Precision-Recall* криві	Агрегація по класах, пороги *IoU* 0.5–0.95 на *COCO*-подібних наборах	*YOLO, Faster R-CNN*
Семантична сегментація	*mIoU, F1/Dice, Foreground Accuracy*	Попіксельне оцінювання, час інференсу на *CamVid/Cityscapes*	*U-Net, HRNet*

Щоб об’єктивно оцінити якість контурів було розроблено метрики, зосереджені на геометричних властивостях границь, а не на об’ємному перекритті.

Boundary IoU (BIoU) є ключовим показником, спрямованим на подолання низької чутливості стандартного *IoU* до помилок границь, особливо для великих об’єктів. Замість того, щоб розглядати всі пікселі маски, *BIoU* обчислює перекриття-над-об’єднанням лише для пікселів, розташованих у граничному регіоні (*Boundary Region*).

Основні переваги *BIoU* полягають у його здатності надавати кількісний градієнт, який безпосередньо покращує якість граничної сегментації, а також у його збалансованій реакції на помилки незалежно від розміру об’єкта. *BIoU* зменшує упередженість, яку демонструє стандартний *IoU* щодо великих структур [16].

Метрики на основі відстаней, як-от *Hausdorff Distance* (*HD*), *Average Symmetric Surface Distance* (*ASSD*) і *Normalized Surface Distance* (*NSD*), забезпечують оцінку геометричної розбіжності між контурами. Синергетичне використання *BIoU* (як стабільного індикатора загальної якості контуру) та *HD* (як індикатора катастрофічних локальних помилок) гарантує повний контроль геометричної точності.

Для досягнення надійності в критичних системах оцінювання точності має бути доповнене оцінюванням впевненості моделі. *UQ* є обов’язковим елементом для валідації надійності та безпеки. Кількісне оцінювання невизначеності дає змогу встановити, наскільки добре ймовірнісні прогнози моделі (впевненість) відповідають її фактичній точності (калібруванню).

Надійність механізмів *UQ* здебільшого залежить від обраної метрики.

Серед ключових метрик калібрування та локалізації помилок можна виокремити такі:

Expected Calibration Error (ECE) – стандартна метрика калібрування – кількісно оцінює узгодженість між оголошеною впевненістю моделі та її фактичною точністю. Адаптація *ECE* для сегментації є необхідною для валідації надійності прогностичних імовірностей;

– *Predictive Entropy* (предиктивна ентропія) значно перевершує інші метрики *UQ*, як-от варіація чи взаємна інформація, з погляду як калібрування (*ECE-label*), так і здатності до локалізації помилок (*Uncertainty-Error overlap*). Висока ентропія в певному регіоні сцени чітко вказує, де саме прогноз є найменш надійним.

– *Area Under the Sparsification Error curve (AUSE)* оцінює ефективність самого механізму *UQ*; вимірює, наскільки успішно оцінки невизначеності можуть бути використані для ідентифікації та відкидання найбільш невпевнених пікселів або патчів, підвищуючи цим загальну надійність системи. Застосування *AUSE* підтверджує, чи є механізм *UQ* практично корисним для розроблення механізмів фільтрації та відмови.

5. Обговорення результатів дослідження

Комплексне оцінювання точності ЗНМ для аналізу статичних сцен вимагає паралельного використання кількох груп метрик. Запропоновано гібридний фреймворк, який забезпечить повну валідацію моделі, покриваючи загальний об’єм, геометричну якість та надійність:

Overlap (*IoU, Dice Score*) – базова оцінка загального покриття.

Boundary / Geometry (*BIoU, HD, ASSD*) – гарантування геометричної точності та виявлення граничних помилок.

Reliability / Calibration (*ECE, Predictive Entropy, AUSE*) – оцінювання впевненості та локалізації помилок.

Цей підхід розв’язує проблему узгодження цілей: моделі, які оптимізуються під функції втрат, що містять гранично-орієнтовані метрики (наприклад, *Generalized Surface Loss*), будуть належним чином винагороджені в процесі валідації за допомогою *BIoU* та *HD*.

BIoU забезпечує більш справедливу й масштабно-збалансовану оцінку продуктивності, що є критично важливим у роботі з гетерогенними сценами, де об’єкти сильно різняться за розміром.

Інтеграція *UQ*, особливо через метрики, які здатні до локалізації помилок, є необхідною для підвищення надійності систем.

Predictive Entropy завдяки своїй високій надійності дає змогу точно ідентифікувати ділянки сцени, де прогноз є найменш впевненим.

Предиктивна ентропія вимірює ступінь "чистоти" розподілу ймовірностей класу в кожному пікселі. Що вища ентропія, то більш невизначений прогноз моделі щодо того, до якого класу належить піксель. Ентропія ефективно використовується для локалізації помилок. Гібридний фреймворк оцінювання – це багатовимірний інструментарій, призначений для забезпечення повної картини продуктивності ЗНМ за межами простого відсотка перекриття. У табл. 3 відтворено взаємодію компонентів запропонованого гібридного фреймворку.

Таблиця 3. *Взаємодія компонентів запропонованого фреймворку*

Метрика	Що вимірює?	Вирішувана проблема	Ключова роль у фреймворку
IoU/Dice	Об’ємне перекриття	Базовий рівень точності	Установлення загального успіху
Boundary IoU (BIoU)	Точність контуру	Нечутливість *IoU* до границь	Геометрична якість, збалансована оцінка
Hausdorff Distance (HD)	Максимальна помилка	Виявлення локальних катастроф	Індикатор безпеки та наявності викидів
Predictive Entropy	Локальна невизначеність	Локалізація ненадійних зон	Забезпечення надійності, механізм відмови
ECE	Калібрування впевненості	Недовіра до ймовірностей	Валідація довіри до прогнозів моделі

Нижче наведений рекомендований протокол для репрезентативного оцінювання моделей аналізу статичних сцен.

1. Чітко задокументувати препроцесинг, аугментації та розмітку; зберегти *seed*-и й опублікувати код.

2. Обирати набір метрик відповідно до завдання: для сегментації – *mIoU + boundary F-score + Pixel Accuracy*; для детекції – *COCO mAP(0.5:0.95) + AP@0.5 + recall@K + analysis by object sizes*.

3. Побудувати довірчі інтервали (*bootstrapping*) для ключових метрик.

4. Провести *cross-dataset evaluation*, якщо мета – узагальнення в змінених доменах.

5. Оцінити калібрування (*ECE, NLL*) і в разі потреби застосувати методи калібрування (*temperature scaling*) або UQ (*ensembles, MC Dropout*).

6. Провести *robustness checks* (шум, змінене освітлення, геометричні трансформації) і, якщо необхідно, *adversarial robustness* аналіз.

Етапи оцінювання, їх основна мета й ключові метрики оцінювання моделей відповідно до рекомендованого протоколу подано в табл. 4.

Таблиця 4. *Етапи оцінювання, їх основна мета й ключові метрики оцінювання моделей*

№	Етап оцінювання	Основна мета	Ключові дії / метрики
1	**Відтворюваність та документація**	Забезпечення можливості повторного отримання результатів	* Документація: препроцесинг, аугментації, розмітка
			* Фіксація всіх *Seed*-ів (*NumPy, ML framework*)
			* Публікація Коду
2	**Релевантний набір метрик**	Об’єктивне оцінювання якості моделі відповідно до задачі	* Сегментація: *mIoU, Boundary F-score, Pixel Accuracy*
			* Детекція: *COCO mAP (0.5:0.95), AP@0.5, Recall@K, AP_S, AP_M, AP_L*
3	**Довірчі інтервали (*Bootstrapping*)**	Оцінювання статистичної значущості та варіативності результатів	* Побудова 95 % довірчих інтервалів для ключових метрик методом ***Bootstrapping***

Кінець таблиці 4

№	Етап оцінювання	Основна мета	Ключові дії / метрики
4	***Cross-Dataset Evaluation***	Оцінювання узагальнення моделі на нові домени	* Тестування моделі на **зовнішньому, *OOD*** *(Out-of-Domain)* наборі даних
5	**Калібрування та *UQ***	Оцінювання надійності передбачуваних імовірностей	* **Оцінювання:** *ECE (Expected Calibration Error), NLL*
			* **Методи (за потреби):** *Temperature Scaling, Ensembles, MC Dropout*
6	***Robustness Checks***	Оцінювання стійкості до природних і ворожих спотворень	* **Природна *Robustness*:** тестування на шум, змінене освітлення, геометричні трансформації
			* ***Adversarial Robustness*:** аналіз стійкості до ворожих атак (напр., *PGD*)

Цей протокол забезпечує чіткий і всебічний підхід до оцінювання моделей аналізу сцен за межами лише метрик точності (*accuracy / mAP*) і беручи до уваги статистичну значущість (*CI*), узагальнення (*cross-dataset*), калібрування (*ECE*) та стійкість (*robustness*).

Результати використання цього протоколу для оцінювання моделі детекції об’єктів *YOLOv8-L* на наборі *COCO* подано в табл. 5.

Таблиця 5. *Результати використання протоколу для оцінювання моделі детекції об’єктів*

№	Етап оцінювання	Деталі реалізації / метрики	Результати	Висновок
1	**Відтворюваність**	Модель: *YOLOv8-L*. Аугментації: *Mosaic, Flip, MixUp, HSV*	Фіксація Global_Seed = 101. Код навчання опубліковано	Висока відтворюваність
2	**Набір метрик**	*COCO mAP (0.5:0.95), AP@0.5, Recall@100*. Аналіз за розмірами.	***mAP:*** 49.8 %. ***AP@0.5*:** 68.2 %. ***AP_S (Small)*:** 31.5 %.	Добра загальна продуктивність, але слабкість до малих об’єктів
3	**Довірчі інтервали**	*Bootstrapping* (500 ітерацій) для *COCO mAP*	95 % *CI* для *mAP*: **[49.1 %, 50.5 %]**	Результат статистично стабільний
4	***Cross-Dataset***	Тестування на ***Open Images V6*** (*OOD*-набір)	*mAP* на *Open Images*: **38.9 %**. (падіння на 10.9 п.п.)	Модель чутлива до **зміни домену** (*Domain Shift*)
5	**Калібрування**	Оцінювання *ECE* та *NLL*. Застосування *Temperature Scaling*	***ECE* до:** 6.2%. **ECE після:** 4.1 %. ***NLL*:** 0.75	Початково перекалібрована, успішно виправлено ***Temperature Scaling***

Кінець таблиці 5

№	Етап оцінювання	Деталі реалізації / метрики	Результати	Висновок
6	***Robustness Checks***	Вплив шуму, зміни освітлення та *adversarial* атак	**Падіння *mAP*:** шуми (-6.7 %), контраст (-8.3 %), ***PGD Attack*** (-48.3 %)	Дуже вразлива до ***Adversarial Attacks*** і зміни контрасту

Приклад використання протоколу оцінювання для завдання семантичної сегментації з використанням моделі *U-Net* на аерофотознімках наведено в табл. 6.

Таблиця 6. *Результати використання протоколу для оцінювання моделі семантичної сегментації*

№	Етап оцінювання	Деталі реалізації / метрики	Результати	Висновок
1	**Відтворюваність**	Модель: *U-Net* (*VGG*-16 як кодувальник). Препроцесинг: нормалізація *RGB* до $[0, 1]$	*Seed: Global_Seed* = 888. Аугментації: випадкове обертання, випадкова зміна масштабу (0.8–1.2)	Забезпечено чітке документування архітектури та аугментацій
2	**Набір метрик**	*mIoU, Boundary F-score, Pixel Accuracy*. Аналіз за класами: вода, рослинність, пісок, будівлі	*mIoU* (загальний): 75.3 %. *Boundary F-score:* 70.1 %. *mIoU* (пісок): 89.5 %. *mIoU* (будівлі): 60.2 %	Хороша загальна продуктивність, але слабкість у сегментації будівель (менші та складніші форми)
3	**Довірчі інтервали**	*Bootstrapping* (700 ітерацій) для метрики *mIoU* (будівлі)	95 % *CI* для *mIoU* (будівлі): [58.5 %, 61.9 %]	Результат для складного класу є статистично значущим
4	***Cross-Dataset***	Тестування на OOD-наборі *Potsdam* (аерофотознімки з іншого міста / камери) з мапінгом класів	*mIoU* на *Potsdam*: 65.9 %. (падіння на 9.4 п.п.)	Модель чутлива до змін у кольоровій гамі та роздільній здатності зображень іншого домену
5	**Калібрування**	Оцінювання *ECE* та *NLL*	*ECE*: 5.1 %. *NLL*: 0.45	Модель має помірне перекалібрування. Необхідне пост-хок калібрування (напр., *Temperature Scaling*)
6	***Robustness Checks***	Тестування на: симуляція хмарності (зміни яскравості та контрасту), стиснення *JPEG* (висока якість)	Падіння *mIoU:* симуляція хмарності (-4.8 %), стиснення *JPEG* (-1.1 %)	Чутлива до затемнення / зміни контрасту, що імітує хмарну погоду

Цей приклад демонструє, як протокол може бути застосований для оцінювання моделі сегментації аерофотознімків, виявляючи її слабкості у сегментації малих об’єктів (наприклад, будівель, зелених насаджень тощо) та чутливість до погодних умов (хмарності).

6. Висновки й перспективи подальших досліджень

Порівняння методик демонструє, що для аналізу статичних сцен оптимально комбінувати метрики: *accuracy* та *F1* для класифікації, *IoU* і *mAP* для детекції. Найефективніша *mAP* для складних сцен і для детекції малих об’єктів.

Перспективи передбачають розроблення гібридних метрик з елементами пояснюваності для кращого розуміння помилок *CNN*.

Дослідження також підтвердило недоліки використання традиційних метрик перекриття (*IoU, Dice Score*) для об’єктивного оцінювання геометричної точності ЗНМ в аналізі статичних сцен через їх нечутливість до граничних помилок (*boundary un-awareness*).

Удосконалення методик оцінювання вимагає переходу до гібридного фреймворку, основаного на двох аспектах:

- гранична точність – упровадження *Boundary IoU (BIoU)* та метрик відстані (*HD, ASSD*) забезпечить необхідну чутливість до помилок на контурах, що є життєво важливим для високоточних геометричних застосувань, таких як *MMDE*;
- надійність – інтеграція кількісного оцінювання невизначеності (*UQ*) з пріоритетом *Predictive Entropy* та *ECE*, є обов’язковою для валідації безпеки та калібрування впевненості моделі.

Запропонований гібридний фреймворк забезпечує більш повну й об’єктивну картину продуктивності ЗНМ, об’єднуючи оцінювання геометричної якості та калібрування впевненості. Упровадження гранично-орієнтованих метрик стимулює розроблення архітектур і функцій втрат, які безпосередньо покращують точність контурів, оскільки ці зусилля будуть належним чином винагороджені. Це підвищує загальну надійність систем комп’ютерного зору в критичних доменах.

Подальші дослідження можуть будуть присвячені адаптації метрик до реального часу й інтеграції з великими моделями, як *Vision Transformers*, розробленню гібридних функцій втрат, що безпосередньо оптимізують гранично-орієнтовані метрики (наприклад, *BIoU Loss* або *Loss*-функції на основі поверхневої відстані), для забезпечення максимальної узгодженості між навчанням і валідацією.

Перспективними також є дослідження та уніфікація *UQ*-фреймворків, зокрема розроблення метрик, які поєднують просторову локалізацію невизначеності з чутливістю до границь (*Boundary UQ*) для підвищення надійності оцінювання критичних об’єктів.

References

1. Aalst, J., Maruccio, F., Simoẽs R., Janssen, T., Wolterink, J., Ooijen, P., Brouwer, Ch. (2025), "Reliability of uncertainty quantification methods for deep learning auto-segmentation in head and neck organs at risk", *Physics in Medicine and Biology*, No. 20. DOI: https://doi.org/10.1088/1361-6560/ae110c
2. Fu, Y., Li, X., Hu, Z. (2021), "Small-Target Complex-Scene Detection Method Based on Information Interworking High-Resolution Network", *Sensors*, No. 21(15). DOI: https://doi.org/10.3390/s21155103
3. Chen, F., Tsou, J. (2022), "Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: An in-depth investigation", *International Journal of Applied Earth Observation and Geoinformation*, No. 112. DOI: https://doi.org/10.1016/j.jag.2022.102865
4. Maxwell, A., Warner, T., Guillén, L. (2021), "Accuracy Assessment in Convolutional Neural Network-Based Deep Learning Remote Sensing Studies – Part 1: Literature Review", *Remote Sensing*, No. 13. DOI: https://doi.org/10.3390/rs13132450
5. Dugăeșescu, A., Florea, A. (2025), "Evaluation and analysis of visual methods for CNN explainability: a novel approach and experimental study", *Neural Computing and Applications*, No. 37, P. 14935–14970. DOI: https://doi.org/10.1007/s00521-025-11282-7
6. Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., Parmar, M. (2024), "A review of convolutional neural networks in computer vision", *Artificial Intelligence Review*, No. 57, P. 99. DOI: https://doi.org/10.1007/s10462-024-10721-6
7. Wang, Ch. (2023), "Calibration in Deep Learning: A Survey of the State-of-the-Art". DOI: https://doi.org/10.48550/arXiv.2308.01222
8. Gawlikowski, J., Tassi, C., et al. (2023), "A survey of uncertainty in deep neural networks. Artificial Intelligence Review". DOI: https://doi.org/10.48550/arXiv.2107.03342
9. Singh, Sh., Yadav, A., Jain, J., Shi, H., Johnson, J., Desai, K. (2024), "Benchmarking Object Detectors with COCO: A New Path Forward". DOI: https://doi.org/10.48550/arXiv.2403.18819
10. Arulananth, T., Kuppusamy, P., Ayyasamy, R., Alhashmi, S., Mahalakshmi, M., Vasanth, K., Chinnasamy, P. (2024), "Semantic segmentation of urban environments: Leveraging U-Net deep learning model for cityscape image analysis", *PLoS ONE*, No 19 (4). DOI: https://doi.org/10.1371/journal.pone.0300767
11. Zhang, J. (2025), "Survey on Monocular Metric Depth Estimation", *Computer Vision and Pattern Recognition*. DOI: https://doi.org/10.48550/arXiv.2501.11841
12. Nemavhola, A., Chibaya, C., Viriri, S. (2025), "A Systematic Review of CNN Architectures, Databases, Performance Metrics, and Applications in Face Recognition", *Information*, No 16 (2), P. 107. DOI: https://doi.org/10.3390/info16020107
13. Classification metrics guide. (2025), "Accuracy vs. precision vs. recall in machine learning: what's the difference?". DOI: https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall
14. Khan, S., Mazhar, T., et al. (2024), "Comparative analysis of deep neural network architectures for renewable energy forecasting: enhancing accuracy with meteorological and time-based features", *Discover Sustainability*, No. 5, P. 533. DOI: https://doi.org/10.1007/s43621-024-00783-5

15. Rayed, M., Islam, S., Niha, S., Jim, J., Kabir, M., Mridha, M. (2024), "Deep learning for medical image segmentation: State-of-the-art advancements and challenges", *Informatics in Medicine Unlocked*, No 47. DOI: https://doi.org/10.1016/j.imu.2024.101504
16. Cheng, B., Girshick, R., Dollár, P., Berg, A., Kirillov, A. (2021), "Boundary IoU: Improving Object-Centric Image Segmentation Evaluation", *EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: https://doi.org/10.1109/CVPR46437.2021.01508

Received (Надійшла) 01.11.2025
Accepted for publication (Прийнята до друку) 09.12.2025
Publication date (Дата публікації) 28.12.2025

Відомості про авторів / About the Authors

Влах-Вигриновська Галина Іванівна – кандидат технічних наук, доцент, Національний університет "Львівська політехніка", доцент кафедри комп'ютеризованих систем автоматики Інституту комп'ютерних технологій, автоматики та метрології, Львів, Україна; e-mail: halyna.i.vlakh-vyhrynovska@lpnu.ua; ORCID ID: https://orcid.org/0000-0003-4429-1578

Кромкач Владислав Олександрович – Національний університет "Львівська політехніка", аспірант кафедри комп'ютеризованих систем автоматики Інституту комп'ютерних технологій, автоматики та метрології, Львів, Україна; e-mail: vlad.kromkach@gmail.com, ORCID ID: https://orcid.org/0009-0001-5608-5715

Vlakh-Vyhrynovska Halyna – PhD (Engineering Sciences), Associate Professor, Lviv Polytechnic National University, Associate Professor at the Department of Computerized Automation Systems, Institute of Computer Technologies, Automation and Metrology, Lviv, Ukraine.

Kromkach Vladyslav – Lviv Polytechnic National University, Postgraduate Student at the Department of Computerized Automation Systems, Institute of Computer Technologies, Automation and Metrology, Lviv, Ukraine.

CHOOSING AN ACCURACY ASSESSMENT METHOD FOR STATIC SCENE ANALYSIS TASKS BASED ON CONVOLUTIONAL NEURAL NETWORKS

The object of the study is the quantitative assessment of the accuracy and reliability of convolutional neural networks predictions in static scene analysis tasks, including semantic segmentation and monocular metric depth estimation. Theoretical, analytical and empirical scientific research **methods** were used: comparative analysis, synthesis, systematization, experimental modeling, and others. **The relevance of the research** is due to the fact that traditional accuracy assessment metrics do not always take into account the specifics of static scene analysis tasks - class imbalance, localization and small objects, noise or variable lighting. This reduces the accuracy assessment of the results and requires the implementation of hybrid metrics, boundary-oriented metrics and uncertainty quantification (UQ) to ensure the reliability and security of the systems. **The aim of the research** is to substantiate

and select the most effective and appropriate accuracy assessment method for static scene analysis tasks based on convolutional neural networks by comparing and systematizing existing metrics, analyzing their advantages and limitations in different classes of tasks, and developing an integrated framework to improve the quality of assessment. To achieve this goal, the following tasks are solved in the article: to conduct a comparative analysis of traditional metrics; to study modern approaches and select relevant metrics and protocols for specific classes of tasks; to develop a conceptual hybrid framework that provides full model validation, taking into account overlap, geometric accuracy and confidence calibration. As a result of the study, the following conclusions were made: for the analysis of static scenes, it is optimal to combine the metrics: *accuracy* and *F1* – for classification, *IoU* and *mAP* – for detection. The most effective are *mAP* for complex scenes and for the detection of small objects. A hybrid framework is proposed that provides full validation of the model, covering the total volume, geometric quality and reliability. This framework combines boundary-oriented metrics to ensure geometric accuracy and the methodology of quantitative uncertainty assessment for confidence calibration and error localization. This solves the problem of the mismatch between the high accuracy of models and the limitations of standard validation metrics. The transition to *Boundary IoU* and distance metrics, in particular *Hausdorff Distance*, will provide scale-balanced and significantly higher sensitivity to errors on the contours, and will serve as a tool for detecting catastrophic local geometric deviations. The conceptual framework stimulates the development of more robust and accurate CNN architectures.

Keywords: computer vision; semantic segmentation; object detection; convolutional neural networks; evaluation metrics.

Бібліографічні описи / Bibliographic descriptions

Влах-Вигриновська Г. І., Кромкач В. О. Вибір методики оцінювання точності для завдань аналізу статичних сцен на основі згорткових нейронних мереж. *Автоматизовані системи управління та прилади автоматики*. 2025. № 4 (187). С. 5–19.
DOI: https://doi.org/10.30837/0135-1710.2025.187.005

Vlakh-Vyhrynovska, H., Kromkach, V. (2025), "Choosing an accuracy assessment method for static scene analysis tasks based on convolutional neural networks", *Management Information Systems and Devises*, No. 4 (187), P. 5–19. DOI: https://doi.org/10.30837/0135-1710.2025.187.005