УДК 004.8

ПОРІВНЯЛЬНИЙ АНАЛІЗ НЕЙРОМЕРЕЖЕВИХ МЕТОДІВ І МОДЕЛЕЙ РОЗПІЗНАВАННЯ ТА СЕГМЕНТАЦІЇ ОБ'ЄКТІВ У ВІЗУАЛЬНИХ ПРОЦЕСАХ БУДІВЕЛЬНОГО СЕРЕДОВИЩА

К. В. Данілов[1,a], В. В. Хайдуров[1]

[1] Навчально-науковий Фізико-технічний інститут

Анотація

Будівельні майданчики характеризуються високою складністю візуальної сцени, що ускладнює автоматизоване виявлення та сегментацію об'єктів. У цьому дослідженні виконано порівняльний аналіз трьох сучасних моделей сегментації екземплярів: Mask R-CNN, YOLOv8l-seg та Mask2Former. Навчання та оцінювання проводили на наборі даних Alberta Construction Image Dataset (ACID) із використанням стандартних метрик COCO (mAP, mAP_{50}, mAP_{75}) та оцінки швидкості інференсу (FPS). Результати показали, що Mask2Former досягає найвищої точності (mAP 79.2%), тоді як YOLOv8l-seg забезпечує найвищу швидкість обробки (33.7 FPS) із високою точністю (mAP 77.1%). Mask R-CNN поступається обом сучаснішим моделям як за точністю, так і за швидкістю. Отримані результати ілюструють компроміс між точністю і продуктивністю в задачах сегментації будівельної техніки та підкреслюють ефективність використання трансформерних архітектур у складних візуальних умовах.

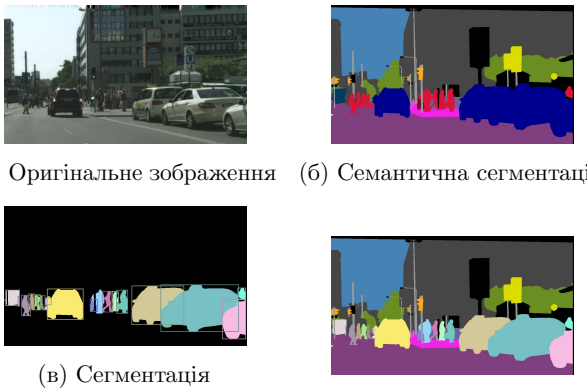
Ключові слова: сегментація екземплярів, комп'ютерний зір, нейронні мережі, будівництво, Mask R-CNN, YOLOv8, Mask2Former, ACID

Вступ

Автоматизація процесів у будівельній галузі є актуальною науково-технічною проблемою, розв'язання якої значною мірою спирається на досягнення в галузі комп'ютерного зору [1, 2]. Будівельні майданчики являють собою динамічні та візуально складні середовища, що містять велику кількість різнорідних об'єктів: персонал, важку техніку, матеріали та конструкції на різних стадіях зведення. Ефективний моніторинг таких середовищ потребує не лише виявлення всіх об'єктів, а й точного визначення їхніх меж та ідентифікації кожного екземпляра.

Однією з ключових задач є сегментація зображень, яка забезпечує деталізоване розуміння сцени на рівні окремих пікселів. Незважаючи на успіхи сегментаційних моделей у загальних наборах даних (COCO, ADE20K), застосування передових методів, зокрема трансформерних архітектур, до зображень будівельних об'єктів залишається недостатньо дослідженим [1]. Крім того, більшість наявних досліджень зосереджуються на обмежених класах об'єктів, що не повністю відображає реальну різноманітність будівельних середовищ.

На рис. 1 наведено приклади різних підходів до сегментації зображень.



(а) Оригінальне зображення (б) Семантична сегментація

(в) Сегментація екземплярів (г) Паноптична сегментація

Рис. 1. Візуальне порівняння типів сегментації [3].

1. Теоретичні основи сегментації екземплярів

Сегментація зображень полягає в розділенні картинки на окремі значущі області, що зазвичай відповідають реальним об'єктам або їх частинам. Формально зображення можна розглядати як функцію $I : \Omega \to \mathbb{R}^k$, де Ω – множина пікселів, а k – кількість каналів (наприклад, $k = 3$ для RGB).

Існує кілька основних підходів до сегментації:

- **Семантична сегментація**: кожен піксель отримує мітку класу з множини $\mathscr{C}$. Об'єкти одного класу не розрізняються між собою (рис. 1(б)).
- **Сегментація екземплярів**: крім класу, виділя-

[a]kosdan-ipt25@lll.kpi.ua

ються окремі об'єкти одного типу. Результатом є множина пар $\{(\boldsymbol{M}_i, c_i)\}_{i=1}^{N}$, де $\boldsymbol{M}_i$ – маска i-го об'єкта, а c_i – його клас. Маски різних об'єктів не перетинаються у межах одного класу (рис. 1(в)).

- **Паноптична сегментація** [3]: комбінує семантичну та сегментацію екземплярів, призначаючи кожному пікселю як клас, так і ідентифікатор: $p : \Omega \to \mathscr{C} \times \mathbb{N}$ (рис. 1(г)).

Основну увагу приділено саме **сегментації екземплярів**, яка дозволяє точно локалізувати окремі об'єкти будівельної техніки навіть за умов перекриття чи часткової видимості.

2. Основні архітектури сегментації екземплярів

2.1. Backbone: поняття та призначення

Backbone – це частина нейронної мережі, яка відповідає за виділення ознак із вхідного зображення $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$. Цей процес можна описати відображенням $\boldsymbol{\phi}_\theta : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H'\times W'\times C'}$, де θ – набір ваг мережі, що навчаються. Отриманий тензор $\boldsymbol{\phi}_\theta(\boldsymbol{I})$ містить компактне, але змістовне представлення сцени, яке далі використовується для детекції та сегментації об'єктів.

Висока якість та інформативність ознак, вилучених компонентом backbone, безпосередньо визначає успішність та точність подальших етапів обробки даних у моделі, включаючи генерацію регіональних пропозицій, класифікацію виявлених об'єктів та прогнозування їхніх сегментаційних масок. Таким чином, вибір сильного та ефективного backbone є необхідною умовою для досягнення високих показників продуктивності загальної архітектури для сегментації екземплярів.

У сучасних архітектурах для екстракції ознак найчастіше використовуються глибокі згорткові мережі або трансформери. Одним із поширених рішень є ResNet-50/101 у поєднанні з Feature Pyramid Networks [4, 5]. Іншим варіантом є архітектура CSPDarknet [6], характерна для моделей YOLO, яка за рахунок розбиття фільтрів та перехресного обміну інформацією дозволяє зменшити обчислювальні витрати без втрати якості. Серед архітектур, заснованих на трансформерах, останнім часом популярності набув ієрархічний Swin Transformer [7]. Він застосовує механізм самоуваги не до всього зображення одразу, а до локальних вікон фіксованого розміру, що робить обчислення більш ефективними для зображень високої роздільної здатності порівняно з класичними трансформерами. При цьому, завдяки механізму зміщення вікон (shifted windows) між послідовними шарами, забезпечується обмін інформацією між сусідніми регіонами, що дозволяє формувати багаторівневе представлення ознак із відносно низькою обчислювальною складністю.

2.2. Mask R-CNN

Mask R-CNN [8] є двоетапною архітектурою для сегментації екземплярів, що розширює Faster R-CNN додаванням окремої гілки для прогнозування масок. Архітектура включає backbone (ResNet-FPN) для екстракції ознак, Region Proposal Network (RPN) для генерації кандидатних регіонів та голови для класифікації, регресії рамок і побудови масок. Точність вирівнювання забезпечується механізмом ROIAlign.

Функція втрат має вигляд:

$$\mathscr{L} = \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{box}} + \mathscr{L}_{\text{mask}},$$

де $\mathscr{L}_{\text{cls}}$ – крос-ентропія класифікації, $\mathscr{L}_{\text{box}}$ – Smooth L1 для регресії координат, а $\mathscr{L}_{\text{mask}}$ – бінарна крос-ентропія для прогнозування маски.

Mask R-CNN забезпечує високу якість сегментації, проте поступається одноетапним моделям за швидкістю інференсу.

2.3. YOLOv8-seg

YOLOv8-seg [9] є представником одноетапних (one-stage) архітектур для одночасної детекції та сегментації екземплярів. Модель використовує високоефективний backbone, зазвичай заснований на архітектурі CSPNet, та anchor-free голови для прогнозування об'єктів.

Навчання моделі базується на мінімізації функції втрат, що складається з декількох компонентів:

$$\mathscr{L} = \mathscr{L}_{\text{box}} + \mathscr{L}_{\text{cls}} + \mathscr{L}_{\text{obj}} + \mathscr{L}_{\text{mask}},$$

де $\mathscr{L}_{\text{box}}$ – втрата регресії рамок, $\mathscr{L}_{\text{cls}}$ – втрата класифікації об'єктів, $\mathscr{L}_{\text{obj}}$ – втрата об'єктності (confidence score), а $\mathscr{L}_{\text{mask}}$ – втрата сегментації, яка зазвичай реалізується як піксельна бінарна крос-ентропія або втрата Dice між передбаченими та істинними масками.

Основною перевагою YOLOv8-seg є висока швидкість інференсу, що робить цю модель особливо придатною для застосувань, де важлива робота в реальному часі.

2.4. Mask2Former

Mask2Former [10] пропонує уніфікований підхід для всіх типів сегментації зображень, використовуючи механізми трансформерів. Архітектура складається з багаторівневого екстрактора ознак (backbone), піксельного декодера та трансформерного декодера з маскованою перехресною увагою (masked cross-attention). Ця архітектура дозволяє генерувати маски безпосередньо, не покладаючись на проміжні етапи, такі як генерація рамок, що є характерним для Mask R-CNN.

Функція втрат у Mask2Former визначається як сума втрати маски та втрати класифікації:

$$\mathscr{L} = \mathscr{L}_{\text{mask}} + \lambda_{\text{cls}}\mathscr{L}_{\text{cls}},$$

де

$$\mathscr{L}_{\text{mask}} = \lambda_{\text{ce}}\mathscr{L}_{\text{ce}} + \lambda_{\text{dice}}\mathscr{L}_{\text{dice}}.$$

Тут $\mathscr{L}_{\text{ce}}$ – бінарна крос-ентропійна втрата, $\mathscr{L}_{\text{dice}}$ – втрата за коефіцієнтом Дайса. Типові

вагові коефіцієнти $\lambda_{ce} = 5.0$, $\lambda_{dice} = 5.0$ та λ_{cls} (зі значенням 2.0 для зіставлених прогнозів та 0.1 для прогнозів без відповідності).

Mask2Former досягає передових результатів за точністю на багатьох бенчмарках сегментації, завдяки ефективному використанню глобального контексту та механізмів уваги, що робить його потужним рішенням для складних задач візуального розуміння сцени.

3. Експериментальна установка

3.1. Набір даних

У дослідженні використано набір даних Alberta Construction Image Dataset (ACID) [2, 11], що містить 10000 зображень будівельної техніки в умовах реальних будівельних майданчиків. Датасет складається з 10 класів будівельних машин, при цьому всі зображення розмічені на рівні екземплярів, що дає можливість використовувати його як для детекції, так і для сегментації об'єктів. Зображення характеризуються значною варіативністю сцен, мають різні умови освітлення, перекриття об'єктів та різноманітні ракурси (рис. 2). Для експериментів датасет було розподілено на тренувальну, валідаційну та тестову вибірки у співвідношенні 80%, 10%, 10% відповідно.

Рис. 2. Приклади зображень з датасету ACID

3.2. Метрики оцінювання

Для оцінки точності сегментації екземплярів використовується метрика Intersection over Union (IoU), яка визначає ступінь перекриття між передбаченою маскою M_{pred} та істинною маскою M_{gt}:

$$\mathrm{IoU}(M_{pred}, M_{gt}) = \frac{|M_{pred} \cap M_{gt}|}{|M_{pred} \cup M_{gt}|}. \quad (1)$$

Передбачення вважається істинно позитивним (True Positive, TP), якщо його IoU з істинною маскою перевищує певний поріг t (напр., $t = 0.5$). Інакше, це хибно позитивне (False Positive, FP) передбачення. Істинна маска, для якої не знайдено відповідного передбачення з IoU $\geq t$, вважається хибно негативною (False Negative, FN).

На основі цих показників розраховуються точність (Precision) та повнота (Recall):

$$\mathrm{Precision} = \frac{TP}{TP + FP},$$
$$\mathrm{Recall} = \frac{TP}{TP + FN}.$$

Змінюючи поріг впевненості моделі, можна побудувати криву Precision-Recall. Середня точність (Average Precision, AP) для одного класу обчислюється як площа під цією кривою:

$$\mathrm{AP} = \int_0^1 P(R)\, dR,$$

де $P(R)$ – значення точності при рівні повноти R.

У стандарті COCO [12] використовується усереднена AP по декількох порогах IoU t. Основні метрики, що використовуються:

- **mAP** (або AP) – основна метрика, усереднена AP по 10 порогах IoU від 0.5 до 0.95 з кроком 0.05 та по всіх класах.
- **mAP$_{50}$** (або AP$_{50}$) – усереднена AP по класах при порозі IoU $t = 0.5$.
- **mAP$_{75}$** (або AP$_{75}$) – усереднена AP по класах при порозі IoU $t = 0.75$.

Також оцінювалася швидкість інференсу (Frames Per Second, FPS).

3.3. Параметри навчання та середовище

Експерименти проведено на графічному процесорі NVIDIA L4 24 ГБ у середовищі Google Colab з використанням бібліотеки PyTorch, Detectron2 та Ultralytics. Обрано стандартні параметри навчання та аугментації (масштабування, обертання, віддзеркалення тощо) з урахуванням розміру зображень. Тривалість кожного навчального сеансу (кількість ітерацій чи епох) добиралася з урахуванням збіжності моделі та наявних обчислювальних ресурсів.

4. Результати та аналіз

У цьому розділі наведено порівняння трьох моделей сегментації за основними кількісними показниками та візуальними результатами на тестовій вибірці.

4.1. Кількісна оцінка моделей

Динаміка зміни функції втрат та mAP під час валідації відображена на рис. 3, а підсумкові кількісні показники моделей наведено в таблиці 1. Mask2Former досягає найвищої середньої точності (mAP) на рівні 79.2%. Втім його швидкість обробки становить лише 5.4 FPS, що суттєво нижче порівняно з YOLOv8l-seg (33.7 FPS) та Mask R-CNN (19.3 FPS).

YOLOv8l-seg демонструє високу продуктивність за швидкістю та конкурентоспроможну точність (77.1% mAP), що робить його придатним для застосувань у реальному часі. Mask R-CNN поступається за всіма основними метриками, відображаючи обмеження класичних підходів у контексті сучасних завдань сегментації.

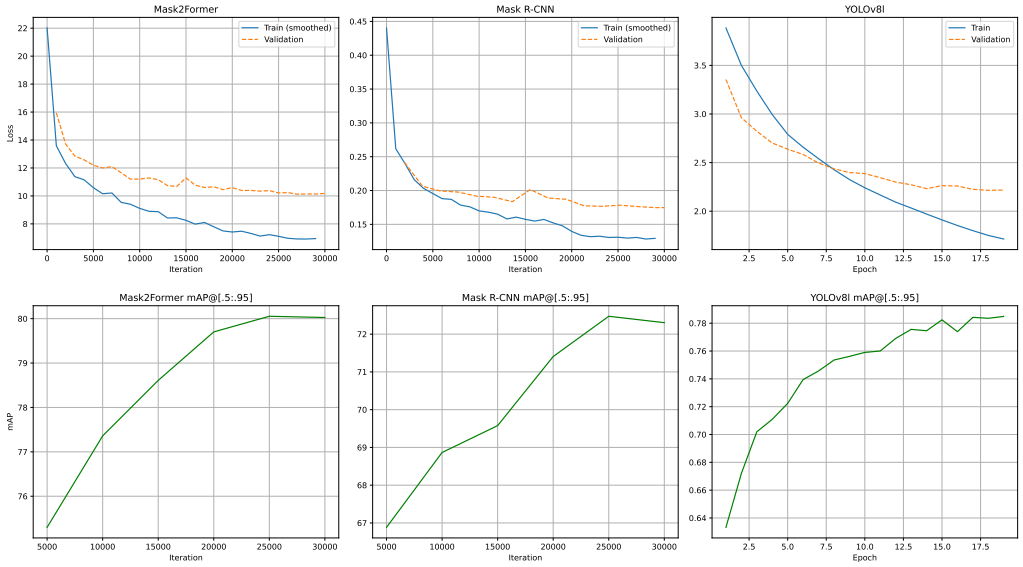


Рис. 3. Зміна функції втрат та mAP@[0.5:0.95] на валідації під час навчання моделей Mask2Former, Mask R-CNN та YOLOv8l

Модель	**Backbone**	**mAP**	**mAP$_{50}$**	**mAP$_{75}$**	**FPS**	**Params (M)**
Mask R-CNN	ResNet-101	71.6	92.2	80.3	19.3	63.28
YOLOv8l-seg	modified CSPDarknet53	77.1	94.4	84.8	**33.7**	45.94
Mask2Former	Swin-t	**79.2**	**94.9**	**85.4**	5.4	47.42

Таблиця 1. Порівняння усереднених метрик сегментації моделей на тестовій вибірці

Клас	**Mask R-CNN**	**YOLOv8l-seg**	**Mask2Former**
Екскаватор-навантажувач	79.08	85.06	**87.22**
Бетонозмішувач	83.84	88.19	**90.34**
Коток	84.14	89.11	**90.40**
Бульдозер	75.08	80.73	**83.95**
Самоскид	71.32	**74.81**	73.41
Екскаватор	68.80	76.08	**77.47**
Грейдер	81.86	84.45	**86.54**
Мобільний кран	63.96	66.22	**75.19**
Баштовий кран	31.72	44.14	**44.31**
Фронтальний навантажувач	76.50	82.62	**83.58**

Таблиця 2. mAP за класами для різних моделей

(а) Оригінальне зображення (б) Mask R-CNN

(в) YOLOv8l-seg (г) Mask2Former

Рис. 4. Візуальне порівняння результатів сегментації на прикладі зображення з тестової вибірки

Детальні результати за класами наведено у таблиці 2. Mask2Former стабільно демонструє вищу точність майже для всіх класів, особливо у випадках об'єктів складної геометрії, таких як *мобільний кран*.

4.2. Візуальне порівняння результатів

На рис. 4 проілюстровано візуальні результати сегментації для тестового зображення. Mask2Former забезпечує точне окреслення меж об'єктів, зокрема стріли та ковша екскаватора. Водночас у складних випадках можуть виникати розділення одного об'єкта на кілька масок через розрив у геометрії (як у випадку екскаватора, де ковш і корпус виявилися сегментованими окремо).

YOLOv8l-seg демонструє загалом правильну локалізацію великих об'єктів та точне відтворення основних контурів, що узгоджується з високими кількісними показниками. Однак у зонах перекриття можуть виникати часткові накладання масок та пропуски об'єктів.

Mask R-CNN формує коректні маски для основних об'єктів сцени, однак межі є менш чіткими, а якість сегментації погіршується у випадках складних перекриттів або об'єктів з тонкими деталями.

Висновки

Було здійснено порівняльний аналіз трьох сучасних моделей сегментації екземплярів (Mask R-CNN, YOLOv8l-seg та Mask2Former) для задач виявлення й сегментації будівельної техніки в складних ві-

зуальних умовах. Експериментальні результати на датасеті ACID дають змогу зробити такі висновки:

- **Mask2Former** демонструє найвищі показники точності сегментації (mAP = 79.2%), особливо у класах об'єктів складної геометрії, таких як мобільні крани. Однак модель має обмеження щодо швидкості обробки (5.4 FPS).
- **YOLOv8l-seg** забезпечує найбільш практичний компроміс між точністю (mAP = 77.1%) та швидкістю обробки (33.7 FPS), що робить його ефективним для систем реального часу. Модель добре сегментує великі об'єкти з чіткими контурами, хоча в складних сценах може втрачати дрібні об'єкти та деталі у зонах перекриття.
- **Mask R-CNN** демонструє нижчі результати серед розглянутих моделей як за точністю, так і за швидкістю обробки. Хоча він стабільно виявляє основні об'єкти сцени, якість масок погіршується у випадках складних перекриттів.

Підсумовуючи, Mask2Former доцільно застосовувати в задачах, де пріоритетом є максимальна точність і детальність сегментації, тоді як YOLOv8l-seg є кращим вибором для систем із обмеженими обчислювальними ресурсами або для застосувань у реальному часі. Mask R-CNN може розглядатися як базове рішення для менш вимогливих завдань сегментації будівельної техніки.

Отримані результати також свідчать про необхідність подальших досліджень із метою підвищення стійкості моделей до перекриттів об'єктів та поліпшення локалізації зі складною структурою.

Перелік використаних джерел

1. Transformer-based automated segmentation of recycling materials for semantic understanding in construction / X. Wang, W. Han, S. Mo, T. Cai, Y. Gong, Y. Li, Z. Zhu // Automation in Construction. — 2023. — Т. 154. — С. 104983.
2. *Xiao B.*, *Kang S.-C.* Development of an image data set of construction machines for deep learning object detection // Journal of Computing in Civil Engineering. — 2021. — Т. 35, № 2. — С. 05020005.
3. Panoptic segmentation / A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollár // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2019. — С. 9404—9413.
4. Deep residual learning for image recognition / K. He, X. Zhang, S. Ren, J. Sun // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2016. — С. 770—778.
5. Feature Pyramid Networks for Object Detection / T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — 07.2017.
6. CSPNet: A new backbone that can enhance learning capability of CNN / C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. — 2020. — С. 390—391.
7. Swin transformer: Hierarchical vision transformer using shifted windows / Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo // Proceedings of the IEEE/CVF international conference on computer vision. — 2021. — С. 10012—10022.
8. Mask R-CNN / K. He, G. Gkioxari, P. Dollár, R. Girshick // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — 2017. — С. 2961—2969.
9. *Jocher G.*, *Qiu J.*, *Chaurasia A.* Ultralytics YOLO. — Вер. 8.0.0. — https://ultralytics.com : Ultralytics, 10.01.2023. — URL: https://github.com/ultralytics/ultralytics.
10. Masked-attention mask transformer for universal image segmentation / B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, R. Girdhar // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2022. — С. 1290—1299.
11. *Xiao B.*, *Wang Y.*, *Kang S.-C.* Deep learning image captioning in construction management: a feasibility study // Journal of Construction Engineering and Management. — 2022. — Т. 148, № 7. — С. 04022049.
12. Microsoft coco: Common objects in context / T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick // Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. — Springer. 2014. — С. 740—755.