



ГІБРИДНА АРХІТЕКТУРА ATTENTION U-NET З ГЛИБИННО-РОЗДІЛЮВАНИМИ ЗГОРТКАМИ ДЛЯ СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ

Розроблено асиметричну архітектуру нейромережі типу U-Net, оптимізовану для функціонування за умов суворих обмежень на апаратні ресурси оперативної та відеопам'яті обчислювальних систем. Наведено схему вибіркової модифікації структурних компонентів, яка передбачає збереження повнорозмірних стандартних згорткових операцій на етапі стиснення просторових ознак в енкодері для максимального вилучення низькорівневих геометричних особливостей та інтеграцію полегшених глибинно-розділюваних згорткових операцій на етапі відновлення роздільної здатності в шарах декодера. Введено блоки адитивної м'якої уваги для проміжних карт ознак у структурі пропущених зв'язків задля спрямованої стабілізації процесу навчання та повної компенсації можливої втрати точності цифрової сегментації зображень через полегшення внутрішньої архітектури шарів декодера. Проведено серію ідентичних експериментальних запусків процесу навчання моделі з варіативним набором початкових випадкових вагомостей шарів для забезпечення високої статистичної достовірності й об'єктивності порівняльного аналізу. Досліджено поведінку запропонованого гібридного підходу до побудови архітектури нейромережі на спеціалізованому біологічному наборі даних низькоконтрастних об'єктів Kvasir-SEG, що містить одну тисячу цифрових зображень із високою текстурною неоднорідністю навколишнього фону. З'ясовано, що запропонована архітектура нейромережі забезпечує середній показник точності за метрикою Dice на рівні 0,819, демонструючи практичну еквівалентність за якістю сегментації меж зображення порівняно зі стандартним прототипом нейромережі U-Net. При цьому загальну кількість навчуваних параметрів нейромережі було успішно зменшено на 14,6 % від базового обсягу. Встановлено трикратне зниження амплітуди коливальних валідаційної метрики під час тренування моделі та зменшення стандартного відхилення результатів до 0,004, що чітко вказує на підвищену стійкість сформованої системи до умов початкової ініціалізації вагомостей шарів. Підтверджено ефективність застосування комбінованого функціоналу втрат на основі попіксельної бінарної крос-ентропії та просторового перекриття масок у пропорції 0,3 до 0,7 для надійного пригнічення хибнопозитивних артефактів на тлі текстурних неоднорідностей. Проаналізовано швидкісні характеристики моделі на графічному пришвидшувачі NVIDIA T4 та доведено збереження низької латентності інференсу на рівні 26,41 мс на одне зображення.

Ключові слова: комп'ютерний зір; штучні нейронні мережі; моделі енкодер-декодер; локалізація меж об'єктів; селективне зважування ознак об'єктів; медичний набір даних Kvasir-SEG; коефіцієнт Дайса-Соренсена; пристрої з обмеженими ресурсами.

Вступ / Introduction

Семантична сегментація цифрових зображень є однією з ключових задач комп'ютерного зору. Точне автоматичне виділення меж цільових об'єктів безпосередньо впливає на якість оброблення даних у системах автономного водіння, автоматизації промислового виробництва, робототехніці та системах моніторингу складних просторових сцен. Історичним переламом у цій галузі стало створення архітектури нейромережі U-Net [9], яка заклала основу симетричних топологій із пропущеними зв'язками (англ. *Skip Connections* – структурні обхідні шляхи для безпосереднього передавання просторових ознак низького рівня з енкодера до декодера).

Розвиток цього напрямку у світовій науковій літературі пов'язаний із працями дослідників, які обґрунтува-

ли концепцію глибинно-розділюваних згорток [5], що є математично полегшеним варіантом звичайної згорткової операції, а також інтегрували в архітектуру нейромережі U-Net блоки адитивної уваги [2].

Наразі у світовій практиці вже успішно реалізовано низку підходів до зниження обчислювальної складності нейромереж через повну або часткову заміну стандартних згорткових блоків на їхні полегшені аналоги. Проте, попри значні результати, у цій предметній області залишається невирішеною фундаментальна суперечність: радикальне полегшення математичних операторів призводить до втрати здатності мережі виявляти складні ознаки об'єктів. Це є критичним під час оброблення зображень у доменах, для яких характерні низький візуальний контраст і висока текстурна неоднорідність фону. Прикладом такого домену може бути медицина. З

© Copyright 2026 by the author(s)

Столярчук Євген Юрійович, аспірант, кафедра програмного забезпечення комп'ютерних систем.

Email: evgenstolyarchuk@gmail.com; <https://orcid.org/0009-0009-8008-9423>

Сулема Євгенія Станіславівна, д-р техн. наук, професор, завідувач кафедри програмного забезпечення комп'ютерних систем.

Email: sulema@pzks.fpm.kpi.ua; <https://orcid.org/0000-0001-7871-9806>

Цитування за ДСТУ: Столярчук Є. Ю., Сулема Є. С. Гібридна архітектура attention U-Net з глибинно-розділюваними згортками для сегментації зображень. Науковий вісник НЛТУ України. 2026, т. 36, № 4. С. 44–50.

Citation APA: Stoliarchuk, I. Yu., & Sulema, Ye. S. (2026). Hybrid attention U-Net architecture with depthwise separable convolutions for image segmentation. *Scientific Bulletin of UNFU*, 36(4), 44–50. <https://doi.org/10.36930/40360405>

іншого боку, спроби компенсувати цю втрату за рахунок розгалужених багаторівневих систем уваги створюють надмірну обчислювальну складність, що нівелює переваги оптимізації.

Отже, актуальним науковим завданням є розроблення методики сегментації зображень, яка забезпечувала б скорочення кількості параметрів та обчислювальних операцій без погіршення якості такої сегментації. Одним із перспективних підходів до вирішення цього завдання є поєднання механізму уваги Attention U-Net із вибірко-вим використанням глибинно-розділюваних згорток у найбільш обчислювально витратних компонентах мережі.

Об'єкт дослідження – процеси цифрової сегментації зображень за допомогою нейронних мереж.

Предмет дослідження – методи і засоби модифікації архітектури нейронної мережі U-Net, які дадуть змогу знизити її обчислювальну складність та кількість параметрів зі збереженням ефективності екстракції ознак.

Мета роботи – зменшити обчислювальну складність нейронної мережі U-Net шляхом розроблення її гібридної архітектури з attention-механізмами та глибинно-розділюваними згортками, що дасть можливість скоротити кількість параметрів моделі без втрати здатності до екстракції складних ознак мережі.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- проаналізувати наявні підходи до зниження обчислювальної складності згорткових моделей та методи інтеграції механізмів уваги, що дасть змогу визначити найменш ефективні та найбільш ресурсомісткі компоненти базової архітектури нейромережі U-Net;
- обґрунтувати структурно-логічну схему модифікованої нейронної мережі, що дасть можливість теоретично поєднати переваги Attention U-Net із концепцією селективного полегшення операторів;
- реалізувати вибіркоче впровадження глибинно-розділюваних згорток у найбільш обчислювально витратні блоки моделі, що забезпечить скорочення кількості параметрів моделі без втрати здатності до екстракції її складних ознак;
- здійснити програмну реалізацію створеної гібридної архітектури нейромережі та тренування її моделі на тестових наборах даних, які допоможуть оцінити точність моделі у складних доменах із низьким візуальним контрастом і неоднорідним фоном;
- виконати комплексний порівняльний аналіз запропонованого рішення із класичними аналогами за показниками швидкодії та обсягу пам'яті, які приведуть до об'єктивного підтвердження обчислювальної ефективності розробленого методу.

Аналіз останніх досліджень та публікацій. У науковій праці [11] автори проаналізували сучасні тенденції сегментації медичних зображень і виокремили низку ключових напрямів розвитку після появи нейромережі U-Net, зокрема – механізми уваги, трансформерні архітектури нейромережі, контрастивне навчання та дистиляцію знань. Стосовно полегшених архітектур нейромереж дослідники зауважили, що висока точність згорткових і трансформерних моделей супроводжується значними обчислювальними витратами, які обмежують їх застосування, та виділили глибинно-розділювані згортки як один із засобів зниження цих витрат, а калібрування ознак на основі уваги – як спосіб розширення рецептивного поля без нарощування кількості параметрів.

У роботі [3] автори систематизували понад 100 модифікацій нейромережі U-Net за шістьма напрямками,

серед яких вдосконалення пропущених зв'язків, зміни структури енкодера, покращення вузького місця, трансформерні розширення, збагачення наведення ознак та ймовірнісні підходи. Дослідники встановили, що найменш витратним є вдосконалення пропущених зв'язків за допомогою блоків адитивної уваги, і водночас зауважили, що заміна стандартних згорток на глибинно-розділювані знижує обчислювальну складність, проте супроводжується ризиком втрати точності екстракції ознак.

У праці [2] автори запропонували інтеграцію блоків адитивної м'якої уваги у структуру пропущених зв'язків архітектури нейромережі U-Net та довели на наборі даних CT-150, що впровадження механізму уваги підвищує середній коефіцієнт Dice з 0,814 до 0,840 за зростання кількості параметрів тільки на 8,8 %. Дослідники обґрунтували, що увага на пропущених зв'язках дає змогу пригнічувати нерелевантні фонові ділянки без залучення зовнішніх моделей локалізації.

У дослідженні [7] автори проаналізували понад 300 публікацій щодо інтеграції механізмів уваги в архітектурі нейромережі комп'ютерного зору та обґрунтували перевагу динамічного зважування ознак у пропущених зв'язках і на стороні декодера для точної локалізації меж об'єктів, оскільки застосування уваги до загальних ознак енкодера може випадково пригнітити важливу структурну інформацію низького рівня.

У роботі [15] автори систематизували підходи до сегментації зображень поліпів за останні п'ять років і виокремили три ключові виклики, зумовлені варіативністю об'єктів за формою, розміром і текстурою, через які цільові структури важко відокремити від навколишніх тканин. Дослідники порівнювали згорткові, трансформерні та гібридні сімейства архітектур нейромереж і з'ясували, що трансформерні моделі забезпечують високу точність ціною значних апаратних ресурсів, тоді як полегшені згорткові архітектури нейромереж зберігають прийнятну ефективність за істотно нижчого обчислювального навантаження.

У роботі [8] автори представили сегментований набір даних Kvasir-SEG, що містить 1000 зображень із бінарними масками поліпів у широкому діапазоні форм, розмірів і текстур, та навели базовий результат порівняльної моделі ResUNet із середнім значенням коефіцієнта Dice 0,788, який слугує орієнтиром для оцінювання нових архітектур нейромереж.

У праці [6] автори представили гібридну мережу MDS U-Net, у якій глибинно-розділювані згортки інтегровано в енкодер, а компенсацію виразності забезпечено подвійною увагою на рівні вузького місця та блоками уваги на пропущених зв'язках, що дало можливість знизити кількість параметрів від 34,5 до 23,5 млн. Дослідники підтвердили, що вилучення блоків уваги знижує точність сегментації зображень, тобто точність полегшеної мережі істотно залежить від компенсаційних механізмів.

У роботі [17] автори описали модель DSCA-Net із суцільною заміною стандартних згорток на глибинно-розділювані в енкодері та декодері, доповнену трьома окремими модулями уваги, що знизило кількість параметрів від 7,76 до 2,2 млн, тобто на 71,6 %. Проте дослідники зазначили, що таке радикальне полегшення потребувало істотного ускладнення компенсуювальних модулів для відтворення якості меж об'єктів.

Аналіз літературних джерел свідчить, що проблема пошуку балансу між обчислювальною складністю та

точністю моделей залишається відкритою. У наявних рішеннях відсутні підходи до побудови асиметричних топологій нейромережі U-Net, які б поєднували повнофункціональний згортковий енкодер із вибірковою впровадженням полегшених шарів глибинно-розділюваних згорток та механізмів уваги на пропущених зв'язках. Це зумовлює потребу нашого дослідження, спрямованого на розроблення гібридної архітектури нейромережі з асиметрично полегшеним декодером, що забезпечує скорочення кількості параметрів без деградації точності.

Матеріали та методи дослідження. Дослідження базуються на модифікації архітектури згорткової нейромережі U-Net. Гібридну модель побудовано шляхом інтеграції механізмів адитивної уваги у пропущені зв'язки та використання глибинно-розділюваних згорток у декодері. Програмну реалізацію та навчання виконано у фреймворку PyTorch на хмарній платформі Google Colab (GPU NVIDIA T4). Валідацію здійснено на наборі даних Kvasir-SEG із застосуванням методів просторової аугментації. Оптимізацію параметрів виконано алгоритмом Adam за комбінованою функцією втрат, а оцінювання обчислювальної ефективності – за метриками коефіцієнта Дайса та затримки інференсу.

Результати дослідження та їх обговорення / Research results and their discussion

В основу запропонованої архітектури нейромережі покладено заміну стандартних згорткових шарів на глибинно-розділювані згортки, концепцію яких було детально обґрунтовано в роботі [5].

Обчислювальна складність стандартної згортки. Розглянемо стандартний згортковий шар, що приймає на вхід карту ознак розміром $D_F^2 \cdot M$, де D_F – просторова ширина та висота, а M – кількість вхідних каналів. На виході згорткового шару з ядром D_F^2 отримуємо карту ознак розміром $D_F^2 \cdot N$, де N – кількість вихідних каналів.

Загальну кількість обчислювальних операцій розраховують за формулою:

$$Cost_{std} = D_K^2 \cdot M \cdot D_F^2 \cdot N. \quad (1)$$

Декомпозиція на глибинну та точкову згортки. Тоді як стандартна згортка одночасно фільтрує та комбінує вхідні дані в новий набір виходів за один крок, глибинно-розділювана згортка розділяє цей процес на два окремі шари.

Перший – це глибинна згортка. Вона виконує роль просторового фільтра, застосовуючи один фільтр до кожного вхідного каналу окремо. Кількість операцій на цьому етапі можна подати такою формулою:

$$Cost_{dw} = D_K^2 \cdot M \cdot D_F^2. \quad (2)$$

Другим кроком є точкова згортка. Вона застосовує фільтр розміром 1×1 для того, щоб об'єднати результати першого етапу. Саме тут відбувається змішування інформації між каналами для створення нових складних ознак. Кількість операцій наведено такою формулою:

$$Cost_{pw} = M \cdot N \cdot D_F^2. \quad (3)$$

Отже, загальна кількість операцій буде відповідати сумі кількості операцій на двох етапах:

$$\begin{aligned} Cost_{dsc} &= Cost_{dw} + Cost_{pw} = \\ &= D_K^2 \cdot M \cdot D_F^2 + M \cdot N \cdot D_F^2 = M \cdot D_F^2 \cdot (D_K^2 + N). \end{aligned} \quad (4)$$

Коефіцієнт пришвидження. Для оцінювання обчислювальної ефективності розробленого методу порівняємо витрати обох методів. Співвідношення складності глибинно-розділюваної згортки до стандартної виглядає так:

$$Cost = \frac{Cost_{dsc}}{Cost_{std}} = \frac{M \cdot D_F^2 \cdot (D_K^2 + N)}{D_K^2 \cdot M \cdot N \cdot D_F^2} = \frac{1}{N} + \frac{1}{D_K^2}. \quad (5)$$

Враховуючи топологію архітектури нейромережі U-Net, де кількість вихідних каналів N на різних рівнях мережі варіюється від 64 до 1024, значення першого доданка стає мізерно малим. Отже, основний вииграш в обчислювальній ефективності визначається розміром ядра згортки. Оскільки в завданнях комп'ютерного зору стандартом є використання ядра розміром 3×3 (тобто $D_K = 3$), це дає змогу стверджувати, що заміна стандартних згорткових блоків на їх розділювані аналоги здатна забезпечити скорочення кількості операцій приблизно у 8-9 разів.

Механізм уваги та його інтеграція в архітектуру нейромережі. Використання глибинно-розділюваних згорток дає змогу істотно полегшити модель, проте таке спрощення має зворотний бік. Через зменшення кількості параметрів мережа стає менш стійкою до візуального шуму, що особливо критично для сегментації зображень зі складними умовами знімання, де шуми та артефакти можуть бути візуально подібними на цільові об'єкти. Для компенсації цього ефекту в архітектуру нейромережі було інтегровано механізм селективної м'якої уваги.

Структура та механізм роботи. Коефіцієнт уваги α – це ваговий множник, що визначає відносну важливість конкретного елемента вхідних даних для подальшого оброблення нейронною мережею. Механізм класифікують як м'яку увагу, оскільки використання сигмоїдальної функції активації забезпечує отримання неперервних вагових коефіцієнтів у діапазоні $[0, 1]$. Це робить архітектуру нейромережі диференційованою, даючи можливість оптимізувати параметри уваги без посередньо під час тренування мережі методом зворотного поширення помилки.

На рис. 1 наведено схему блока адитивної м'якої уваги, інтегрованого у структуру пропущених зв'язків нейромережі. Блок адитивної м'якої уваги (англ. *Additive Soft Attention*), також відомий як увага Багданау (англ. *Bahdanau Attention*), обчислює ступінь відповідності між вектором стану декодера (s_t) та векторами станів енкодера (h_i) через операцію додавання після їхнього лінійного перетворення. Основним завданням цього модуля є обчислення коефіцієнтів релевантності α , що дають змогу моделі фокусуватися на цільових об'єктах сегментації.

Блок приймає два тензори – вектор ознак x_t від енкодера, що містить просторові деталі, та гейтувальний сигнал g з глибшого рівня декодера, що містить глобальний контекст. Обидва сигнали проходять крізь незалежні згорткові шари W_g та W_x з ядром 1×1 . Це перетворення приводить вхідні дані до єдиного проміжного простору ознак F_{int} без зміни їхньої топології. Карти сумуються, підсилюючи спільні ознаки, та пропускаються крізь функцію активації ReLU, яка забезпечує нелінійність та відсікає негативні значення. Після додаткової згортки ω сигнал проходить крізь Sigmoid-активацію, яка формує коефіцієнт уваги $\alpha \in [0, 1]$. Отримана маска вказує на ймовірність належності кожного піксе-

ля до об'єкта інтересу. Оскільки сигнал g має меншу роздільну здатність, карта α масштабується до розміру вхідного тензора x_l . Фінальна операція – попіксельне множення вхідних ознак на коефіцієнт α .

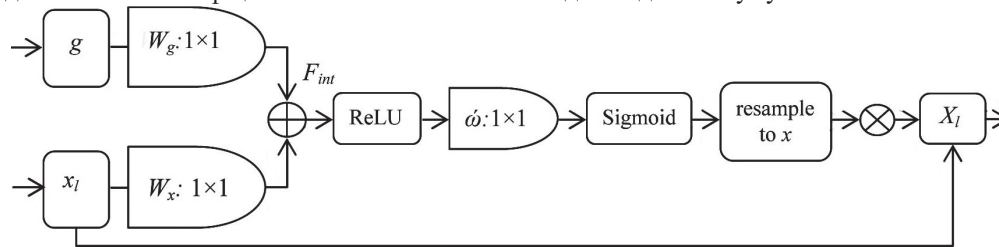


Рис. 1. Схема блока адитивної м'якої уваги / Schematic of the additive soft attention block

Аналіз обчислювальної ефективності. Головною метрикою оцінювання якості сегментації зображень є коефіцієнт Дайса-Соренсена [4, 16], що вимірює перекриття між передбаченою маскою A та еталонною маскою B , який визначають такою логічною формулою:

$$DSC(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|}. \quad (6)$$

В оригінальному дослідженні [2] проведено детальне порівняння базової архітектури нейромережі U-Net та її модифікації з інтегрованими блоками уваги на спеціалізованому наборі даних CT-150. Набір даних CT-150 містить 150 тривимірних сканів і є репрезентативним бенчмарком для задач локалізації структур зі складною морфологією, значною варіативністю форми та низьким контрастом відносно навколишнього фону. Результати експериментів на цьому наборі даних підтверджують, що впровадження механізму м'якої уваги забезпечує стабільний приріст точності сегментації зображень за незначних обчислювальних витрат. Аналіз отриманих показників свідчить, що спостерігається зростання середнього DSC від $0,814^{\pm 0,116}$ до $0,840^{\pm 0,087}$. При цьому обчислювальна вартість залишається низькою, оскільки кількість параметрів моделі зростає тільки на 8,8 %, а час інференсу – на 7,2 %. Така динаміка підтверджує, що механізм м'якої уваги діє як ефективний просторовий фільтр, який завдяки своїй диференційованості дає змогу системі фокусувати ресурси на найбільш інформативних ділянках зображення.

Гібридна архітектура моделі. Запропонована гібридна архітектура нейромережі поєднує принципи Attention U-Net [2] з механізмом глибинно-розділюваних згорток. Ключовою інновацією є асиметричний підхід до побудови мережі: енкодер зберігає стандартні подвійні згорткові блоки для максимально повного вилучення просторових ознак, тоді як декодер використовує блоки глибинно-розділюваних згорток. На всіх пропущених зв'язках розміщені блоки адитивної м'якої уваги, ідентичні оригінальній Attention U-Net.

Така асиметрія обґрунтована функціональними відмінностями між частинами мережі. Енкодер відповідає за формування багатого наведення ознак і потребує повної виразності стандартних згорток. Декодер натомість переважно виконує збільшення роздільної здатності та уточнення меж – задачі, де полегшені блоки втрачають мінімум якості. Це дає змогу скоротити загальну кількість параметрів без істотного впливу на точність сегментації зображень.

Тестовий набір даних. Для тестування гібридної архітектури нейромережі використовувався набір даних Kvasir-SEG, що містить 1000 зображень складних бі-

Результатом роботи блока адитивної м'якої уваги є селективно підсилена карта ознак X_l , у якій вагомість інформативного сигналу максимізована, а вплив фону зведений до мінімуму.

ологічних структур з бінарними масками цільових об'єктів у форматі RGB. Розподіл: 60 % тренування моделі, 20 % валідація і 20 % тестування.

Налаштування параметрів тренування моделі. Обидві моделі навчалися в ідентичних умовах на платформі Google Colab з графічним пришивдшувачем NVIDIA T4. Вхідні зображення масштабувалися до роздільної здатності 256×256 пікселів, розмір батчу становив 8 зображень.

Для оптимізації використовувався алгоритм Adam із початковою швидкістю навчання $1 \cdot 10^{-4}$. Цей алгоритм адаптує швидкість навчання окремо для кожного параметра мережі на основі історії градієнтів, що забезпечує стабільніше навчання порівняно зі звичайним градієнтним спуском. Додатково застосовувався планувальник, що відстежує валідаційну метрику: якщо впродовж 8 епох поспіль покращення не відбувається, швидкість навчання зменшують вдвічі. Це дає змогу моделі точніше досліджувати простір параметрів поблизу знайденого мінімуму. Мінімально допустиме значення швидкості навчання – $1 \cdot 10^{-6}$. Навчання тривало 60 епох, збережена модель відповідає епосі з найвищим значенням Dice Score на валідаційній вибірці.

Як функція втрат використовувалась комбінація бінарної крос-ентропії та Dice Loss у співвідношенні 0,3 до 0,7. Бінарна крос-ентропія оцінює правильність передбачення для кожного пікселя незалежно, проте схильна до домінування фоновому класу за дисбалансу розмірів об'єкта та фону. Dice Loss безпосередньо вимірює перекриття між передбаченою та еталонною масками і нечутливий до цього дисбалансу. Підвищена вагомість Дісе-компоненти змушує модель правильно локалізувати об'єкт як ціле, а не тільки вгадувати фон.

Для підвищення узагальнювальної здатності застосовувалась аугментація, а саме – горизонтальне та вертикальне відзеркалення, повороти на 0° , 90° , 180° та 270° , випадкові зміни яскравості та контрасту.

Аналіз отриманих результатів. Для забезпечення статистичної достовірності кожену модель навчали п'ять разів, щоразу змінюючи початкові випадкові ваги мережі. Це дає змогу оцінити, наскільки стабільний результат незалежно від умов ініціалізації. Результати навчання моделей за кожним запуском наведено в табл. 1.

Табл. 1. Коефіцієнт Dice за окремими запусками моделей / Dice coefficient for individual runs

Запуск моделі	U-Net	Hybrid Attention U-Net
1	0,836	0,827
2	0,829	0,820
3	0,807	0,821
4	0,803	0,812
5	0,815	0,815

Зведені результати для обох моделей наведено в табл. 2 та рис. 2, згідно з якими обидві моделі істотно перевищують результат моделі ResUNet – порівняльної моделі оригінальної статті Kvasir-SEG [8], де середнє значення Dісе коефіцієнта становило 0,788. Різниця між середніми значеннями Dісе коефіцієнта становить менше 0,001, що свідчить про практичну еквівалентність моделей за точністю сегментації зображень в разі скорочення кількості параметрів на 14,6 %.

Табл. 2. Зведене порівняння моделей /
Summary comparison of the models

Модель мережі	Параметри, М	Затримка, мс/зобр.	Коефіцієнт Dісе (середнє $\pm$ стандартне відхилення)
U-Net	23,13	26,77	0,818 $\pm$ 0,014
Hybrid Attention U-Net	19,75	26,41	0,819 $\pm$ 0,004

Стандартне відхилення базової нейромережі U-Net втричі вище (0,014 проти 0,004). Його результат істотно залежить від того, з яких випадкових вагомостей шарів розпочинається навчання. Hybrid Attention U-Net поводить передбачувано в усіх п'яти запусках, що є практичною перевагою в разі розгортання у клінічних умовах: система дає стабільний результат незалежно від умов запуску.

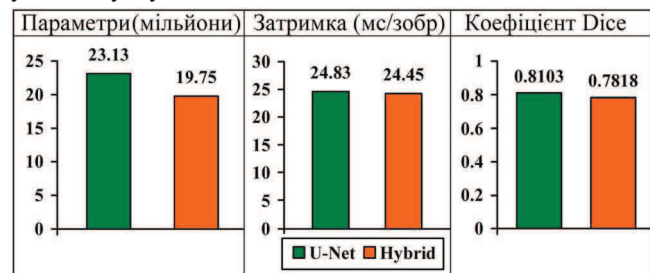


Рис. 2. Порівняння нейромережі U-Net та гібридної архітектури нейромережі на прикладі одного із запусків за трьома показниками: кількість параметрів, латентність та DSC на вибірці Kvasir-SEG / Comparison of U-Net and the hybrid architecture based on a single run across three metrics: number of parameters, latency, and DSC on the Kvasir-SEG dataset

Тривалість сегментації зображень при цьому знижується тільки на 1,4 % – від 26,77 до 26,41 мс на зображення за вимірювання на NVIDIA T4. На потужному серверному пришивидшувачі декодер не є вузьким місцем, тому скорочення параметрів не дає відчутного приросту в швидкості. Проте менша модель споживає менше відеопам'яті, що є критичним в разі розгортання на пристроях з обмеженим VRAM.

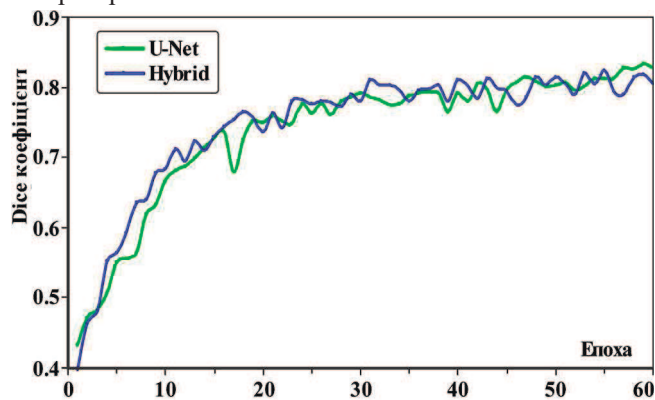


Рис. 3. Криві навчання нейромережі U-Net та гібридної моделі на наборі даних Kvasir-SEG на одному із запусків / Learning curves of U-Net and the hybrid model on the Kvasir-SEG dataset for a single run

Криві навчання обох моделей (рис. 3) демонструють подібну динаміку збіжності, проте гібридна архітектура нейромережі має меншу амплітуду коливань валідаційної метрики, що узгоджується з нижчим стандартним відхиленням у табл. 2.

Варто зазначити роль функції втрат у стабільності очікуваних результатів. Значна вагомість Dісе-компоненти істотно зменшує появу хибнопозитивних артефактів: при рівних вагах 0,5 спостерігалася розпорошена активація у текстурно схожих регіонах. Постоброблення у вигляді відбору найбільшої зв'язної компоненти з порогом 0,6 додатково очищує вивід без потреби в перенавчанні.

Обговорення результатів дослідження. У праці [6] застосовано протилежну до запропонованої стратегію: глибинно-розділювані згортки інтегровано в енкодер, а компенсацію забезпечено важкими модулями уваги, унаслідок чого кількість параметрів знижено від 34,5 до 23,5 млн. Варто зауважити, що аналіз авторів свідчить про повернення кількості параметрів від 18,6 до 23,5 млн саме через модулі уваги, тобто значну частину початкової економії було нівельовано. Оскільки запропоновану в роботі [6] модель валідовано на інших датасетах, пряме кількісне зіставлення з отриманими результатами є некоректним і доцільне тільки на рівні архітектур нейромережної стратегії. Запропонована асиметрична топологія натомість зберігає повнорозмірний енкодер і досягає меншої кількості параметрів в 19,75 млн без залучення важких компенсаційних блоків.

У роботі [17] реалізовано модель DSCA-Net зі скороченням кількості параметрів на 71,6 % відносно базової нейромережі U-Net, шляхом суцільної заміни згорток на глибинно-розділювані та застосування трьох окремих модулів уваги. Це підтверджує, що радикальне полегшення моделі досяжне, проте потребує ускладнення допоміжних механізмів. На відміну від цього, запропонована архітектура нейромережі зберігає простоту її будови за помірного скорочення кількості параметрів на 14,6 %, свідомо надаючи пріоритет стабільності точності її роботи.

У дослідженні [10] наведено мережу EARDS, у якій глибинно-розділювані згортки застосовано винятково в декодері у поєднанні з претренованим енкодером EfficientNet-b0 та блоками уваги на пропущених зв'язках. Збіг ключового архітектурного рішення – полегшення саме декодера зі збереженням повноцінного енкодера – безпосередньо підтверджує обґрунтованість запропонованої архітектури нейромережі. Відмінність полягає в тому, що в роботі [10] використано претренований енкодер, тоді як у цій роботі стандартний згортковий енкодер навчається без попереднього тренування моделі, що робить запропоновану архітектуру нейромережі самодостатньою за відсутності зовнішніх претренованих вагомостей шарів.

Оскільки у праці [12] апробацію проведено на тому самому наборі даних Kvasir-SEG, її результати становлять безпосередній інтерес для порівняння. Автори цієї роботи запропонували полегшену мережу DeepNeXt на основі багатомасштабної уваги та глибинно-розділюваних згорток, що містить тільки 1,51 млн параметрів. Проте пряме зіставлення за коефіцієнтом Dісе є некоректним, оскільки DeepNeXt побудовано не на архітектурі нейромережі U-Net, а на основі DeepLabV3+, а також навчено за іншого співвідношення тренувальної,

валідаційної та тестової вибірок і за меншого розміру вхідних зображень. У концептуальному плані робота [12] демонструє можливість побудови вкрай компактних моделей, тоді як запропонована архітектура нейромережі має іншу мету – зберегти точність на рівні повнорозмірної нейромережі U-Net за помірного скорочення кількості параметрів.

У роботі [13] автори інтегрували в архітектуру нейромережі MedSAM модулі вейвлет-перетворення з каналною увагою та глибинно-розділювану згортку (WCEDSAM), скоротивши кількість параметрів приблизно на 95 % до 15,38 млн, проте зафіксували затримку інференсу на рівні близько 300 мс на зображення. Це підтверджує закономірність, за якою агресивне скорочення параметрів не гарантує зниження латентності, тоді як запропонована модель зберігає низьку затримку 26,41 мс на зображення.

У дослідженні [14] автори реалізували модифіковану нейромережу U-Net для розгортання на пристрої Intel Movidius NCS-2 та скоротили кількість параметрів від 30 до 0,49 млн, підтвердивши інференс безпосередньо на малопотужному апаратному пришвидшувачі. Ця робота підкріплює практичну мотивацію проведеного дослідження щодо придатності полегшених нейромереж типу U-Net для платформ з обмеженими ресурсами.

Отже, зіставлення отриманих результатів із наявними в літературі підтверджує доцільність проведеного дослідження. На відміну від робіт, де полегшення досягається суцільною заміною згорток ціною ускладнення допоміжних модулів [17] або полегшенням енкодера з компенсацією через важку увагу [6], а найближчий за задумом підхід [10] спирається на зовнішній претренований енкодер, запропонована асиметрична топологія забезпечує скорочення кількості параметрів на 14,6 % за практично незмінної точності сегментації зображень та підвищеної стабільності навчання без залучення претренованих вагомостей шарів, чого в проаналізованих джерелах у такому поєднанні досягнуто не було.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – удосконалено архітектуру згорткової нейронної мережі Attention U-Net для семантичної сегментації зображень, яка, на відміну від наявних, містить асиметричну інтеграцію глибинно-розділюваних згорток у блок декодера за збереження повнорозмірних математичних операцій в енкодері та механізмів адитивної м'якої уваги у пропущених зв'язках, що дає змогу скоротити загальну кількість параметрів моделі на 14,6 % без деградації точності сегментації зображень.

Практична значущість результатів дослідження – можна застосувати під час розроблення та розгортання систем комп'ютерного зору для аналізу складних об'єктів на пристроях з обмеженими ресурсами, забезпечуючи при цьому високу точність виділення цільових зон без потреби залучення високопродуктивних серверних графічних пришвидшувачів.

Висновки / Conclusions

Зменшено обчислювальну складність нейронної мережі U-Net шляхом розроблення її гібридної архітектури з attention-механізмами та глибинно-розділюваними згортками, що дало можливість скоротити кількість па-

раметрів моделі без втрати здатності до екстракції складних ознак мережі. За результатами проведеного дослідження можна зробити такі основні висновки.

1. Запропоновано гібридну архітектуру нейромережі на основі нейромережі U-Net, у якій для збереження повноцінної екстракції ознак стандартні згорткові блоки залишено в енкодері, а для зниження її обчислювальної складності в декодері інтегровано глибинно-розділювані згортки.
2. Застосовано блоки адитивної м'якої уваги на всіх пропущених зв'язках моделі задля компенсації обмеженої виразності полегшеного декодера, що, на відміну від повної заміни згорткових операцій, зберегло простоту архітектури нейромережі.
3. Підтверджено ефективність розробленої архітектури нейромережі під час проведення спостережень експерименту на наборі даних Kvasir-SEG: гібридна модель забезпечила середній коефіцієнт Dice на рівні 0,819 (що еквівалентно результату базової нейромережі U-Net – 0,818) за одночасного скорочення кількості параметрів мережі на 14,6 % (від 23,13 млн до 19,75 млн).
4. Встановлено, що розроблена архітектура нейромережі демонструє вищу стабільність процесу навчання порівняно з базовою моделлю, оскільки стандартне відхилення результатів між окремими запусками зменшилося від 0,014 до 0,004.
5. Обґрунтовано доцільність впровадження розробленої моделі гібридної нейромережі для побудови ресурсоефективних систем аналізу складних зображень (зокрема медичних) на апаратних платформах із жорсткими обмеженнями пам'яті та обчислювальної потужності.
6. Окреслено перспективи подальших етапів дослідження, які полягають в оцінюванні ефективності моделі на інших типах даних і мобільних платформах, тестуванні альтернативних схем розміщення полегшених згорток, а також у дослідженні можливостей інтеграції архітектури нейромережі із сучасними трансформерними механізмами.

Заява про використання інструментів штучного інтелекту. Під час підготовки цього рукопису автори застосовували інструменти штучного інтелекту та програмне забезпечення для оброблення природної мови (ChatGPT, Grammarly) винятково задля граматичного редагування, коригування орфографії та стилістичної адаптації складних синтаксичних конструкцій. Автори здійснювали критичний аналіз і подальше редагування згенерованого тексту та несуть повну відповідальність за достовірність, науковий зміст і остаточну редакцію публікації.

References

1. Ali, O., Ali, H., Shah, S. A. A., & Shahzad, A. (2022). Implementation of a Modified U-Net for Medical Image Segmentation on Edge Devices. *IEEE Transactions on Circuits and Systems II*, 69(11), 4593–4597. <https://doi.org/10.1109/TCSII.2022.3181132>
2. Azad, R., Aghdam, E. K., Rauland, A., Jia, Y., Avval, A. H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J. P., Adeli, E., & Merhof, D. (2022). Medical Image Segmentation Review: The Success of U-Net. arXiv:2211.14830. <https://doi.org/10.48550/arXiv.2211.14830>
3. Conze, P.-H., Andrade-Miranda, G., Singh, V. K., Jaouen, V., & Visvikis, D. (2023). Current and Emerging Trends in Medical Image Segmentation With Deep Learning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 7(6), 545–569. <https://doi.org/10.1109/TRPMS.2023.3265863>
4. Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302. <https://doi.org/10.2307/1932409>

5. Howard, A. G., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861. <https://doi.org/10.48550/arXiv.1704.04861>
6. Jha, D., Smedsrud, P. H., Riegler, M. A., Halvorsen, P., de Lange, T., Johansen, D., & Johansen, H. D. (2020). Kvasir-SEG: A Segmented Polyp Dataset. In *International Conference on Multimedia Modeling*, (pp. 451–462). Springer. <https://doi.org/10.48550/arXiv.1911.07069>
7. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., Glocker, B., & Rueckert, D. (2018). Attention U-Net: Learning where to look for the pancreas. arXiv:1804.03999. <https://doi.org/10.48550/arXiv.1804.03999>
8. Qayoom, A., Xie, J., & Ali, H. (2025). Polyp segmentation in medical imaging: challenges, approaches and future directions. *Artificial Intelligence Review*, 58, article number 169. <https://doi.org/10.1007/s10462-025-11173-2>
9. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer International Publishing. <https://doi.org/10.48550/arXiv.1505.04597>
10. Shan, T., Yan, J., Cui, X., & Xie, L. (2023). DSCA-Net: A depthwise separable convolutional neural network with attention mechanism for medical image segmentation. *Mathematical Biosciences and Engineering*, 20(1), 365–382. <https://doi.org/10.3934/mbe.2023017>
11. Sifre, L., & Mallat, S. (2014). *Rigid-motion scattering for image classification*. Ph.D. thesis. Conference paper. <https://doi.org/10.48550/arXiv.1403.1687>
12. Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. *Kongelige Danske Videnskabernes Selskab*, 5(4), 1–34. URL: <https://search.worldcat.org/title/4713331>
13. Wang, C., Wang, S., Shao, S., & Zhai, J. (2024). DeepNeXt: a lightweight polyp segmentation algorithm based on multi-scale attention. *Quantitative Imaging in Medicine and Surgery*. <https://doi.org/10.21037/qims-24-985>
14. Wang, L., Wang, T., Liu, S., et al. (2026). WCEDSAM: A Lightweight Multi-Scale Colonoscopy Polyp-Segmentation Network. *Biology*, 15(9), article ID 707. <https://doi.org/10.3390/biology15090707>
15. Xie, Y., Yang, B., Guan, Q., Zhang, J., Wu, Q., & Xia, Y. (2023). Attention Mechanisms in Medical Image Segmentation: A Survey. arXiv:2305.17937. <https://doi.org/10.48550/arXiv.2305.17937>
16. Zhou, W., Ji, J., Jiang, Y., Wang, J., Qi, Q., & Yi, Y. (2023). EARDS: EfficientNet and Attention-based Residual Depth-wise Separable Convolution for Joint OD and OC Segmentation. *Frontiers in Neuroscience*, 17, article ID 1139181. <https://doi.org/10.3389/fnins.2023.1139181>
17. Zhou, Y., Kang, X., Ren, F., Lu, H., Nakagawa, S., & Shan, X. (2023). A Multi-attention and Depthwise Separable Convolution Network for Medical Image Segmentation. *Neurocomputing*, 564, article ID 126970. <https://doi.org/10.1016/j.neucom.2023.126970>

I. Yu. Stoliarchuk, Ye. S. Sulema

National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine

HYBRID ATTENTION U-NET ARCHITECTURE WITH DEPTHWISE SEPARABLE CONVOLUTIONS FOR IMAGE SEGMENTATION

An asymmetric U-Net neural network architecture has been developed, optimized for operation under severe constraints on hardware resources, specifically the RAM and video memory of computing systems. A scheme for the selective modification of structural components is presented, which involves preserving full-scale standard convolutional operations at the stage of spatial feature compression in the encoder to maximize the extraction of low-level geometric features, and integrating lightweight depthwise separable convolutional operations at the resolution restoration stage in the decoder layers. Additive soft attention blocks have been introduced for intermediate feature maps within the skip connections structure to directionally stabilize the training process and fully compensate for the potential loss of digital segmentation accuracy caused by the lightening of the internal architecture of the decoder layers. A series of identical experimental runs of the model training process with a variable set of initial random weights was conducted to ensure high statistical reliability and objectivity of the comparative analysis. The behavior of the proposed hybrid approach was investigated on the specialized biological dataset of low-contrast objects, Kvasir-SEG, which contains one thousand digital images with high textural heterogeneity of the surrounding background. It was found that the proposed architecture provides an average accuracy score based on the Dice metric at the level of 0,819, demonstrating practical equivalence in boundary segmentation quality compared to the standard U-Net prototype. At the same time, the total number of trainable neural network parameters was successfully reduced by 14,6 % from the baseline volume. A threefold decrease in the amplitude of validation metric fluctuations during training and a reduction in the standard deviation of the results to 0,004 were established, which clearly indicates the increased robustness of the developed system to the conditions of initial layer weight initialization. The effectiveness of applying a combined loss function based on pixel-wise binary cross-entropy and spatial mask overlap in a ratio of 0,3 to 0,7 for reliable suppression of false-positive artifacts against the background of textural heterogeneities was confirmed. The speed characteristics of the model on the NVIDIA T4 graphics accelerator were analyzed, and the preservation of low inference latency at the level of 26,41 ms per image was proven.

Keywords: computer vision, artificial neural networks; encoder-decoder models; object localization; selective feature weighting; Kvasir-SEG; Dice-Sørensen coefficient; resource-constrained devices.