УДК 004.932:004.93

І. М. Кривошей
О. Ю. Софина

АСИМЕТРИЧНА АРХІТЕКТУРА СЕГМЕНТАЦІЇ ЗОБРАЖЕНЬ НА БАЗІ ВІЗУАЛЬНО-МОВНИХ МОДЕЛЕЙ (QWEN3-VL) ТА SAM 3

Вінницький національний технічний університет

Анотація

У роботі розглядається реалізація автоматизованої системи семантичної сегментації зображень з використанням відкритого словника понять. Запропоновано асиметричну архітектуру, що поєднує візуально-мовну модель Qwen3-VL для ідентифікації об'єктів та Segment Anything Model 3 (SAM 3) для їх точної полігональної сегментації. Розроблено інтерактивний веб-додаток на базі Streamlit для просторової аналітики та телеметрії в реальному часі.

Ключові слова: Семантична сегментація, Vision-Language Models, Qwen3-VL, SAM 3, Streamlit, комп'ютерний зір, vLLM..

Abstract

This paper discusses the implementation of an automated semantic image segmentation system using an open vocabulary of concepts. An asymmetric architecture is proposed, combining the Qwen3-VL vision-language model for object identification and Segment Anything Model 3 (SAM 3) for their precise polygonal segmentation. An interactive web application based on Streamlit was developed for real-time spatial analytics and telemetry.

Keywords: Image segmentation, Vision-Language Models, Qwen3-VL, SAM 3, Streamlit, computer vision, vLLM.

Вступ

Сегментація зображень є ключовим етапом у розвитку систем комп'ютерного зору. Сучасні виклики вимагають переходу від класичних алгоритмів та моделей з фіксованим набором класів до інтелектуальних систем, здатних розпізнавати об'єкти за текстовим описом (open-vocabulary). Використання великих візуально-мовних моделей (VLM) у поєднанні з фундаментальними моделями просторового розпізнавання відкриває нові можливості для повної автоматизації візуального аналізу без необхідності попереднього ручного навчання на специфічних датасетах.

Основна частина

У межах дослідження розроблено програмний комплекс, що базується на парадигмі асиметричної перцепції (SOTA AsymmetricPerceptionAdapter). Процес обробки зображення концептуально розділено на дві взаємопов'язані фази.

На першій фазі («Спостерігач») використовується квантована 4-бітна візуально-мовна модель Qwen3-VL-8B-Instruct. Вона аналізує вхідне зображення та генерує масив текстових концептів - коротких описів фізичних об'єктів, присутніх у сцені. Для забезпечення максимальної швидкодії інференсу цієї моделі застосовано інфраструктуру vLLM.

На другій фазі («Екстрактор») згенеровані текстові концепти передаються до фундаментальної моделі сегментації SAM 3 (Segment Anything Model 3) від Facebook. Використовуючи бібліотеку Hugging Face Transformers, SAM 3 обробляє концепти як промпти та формує високоточні полігональні маски для кожного ідентифікованого об'єкта.

Програмне забезпечення реалізовано мовою Python з використанням бібліотек OpenCV, Supervision та NumPy для математичної обробки контурів та масивів даних. Апаратна архітектура передбачає розподіл навантаження: Qwen3-VL функціонує на графічному процесорі (GPU 0), тоді як генерація масок за допомогою SAM 3 виконується на іншому (GPU 1).

Графічний інтерфейс (VisionAI Telemetry Dashboard) створено за допомогою фреймворку Streamlit. Він інтегрує функції завантаження зображень, налаштування порогів чутливості (Confidence threshold) та товщини контурів, а також забезпечує інтерактивну візуалізацію результатів. Блок-схема на рисунку 1 демонструє принцип роботи програмного комплексу.

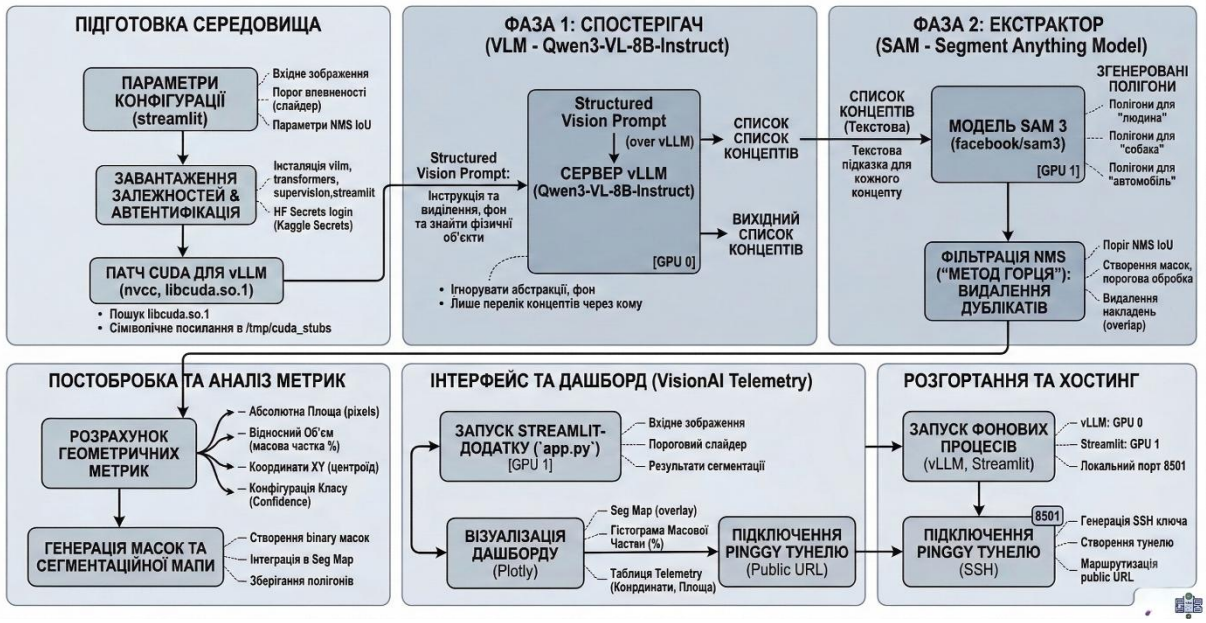


Рисунок 1 – Блок-схема пайплайну сегментації

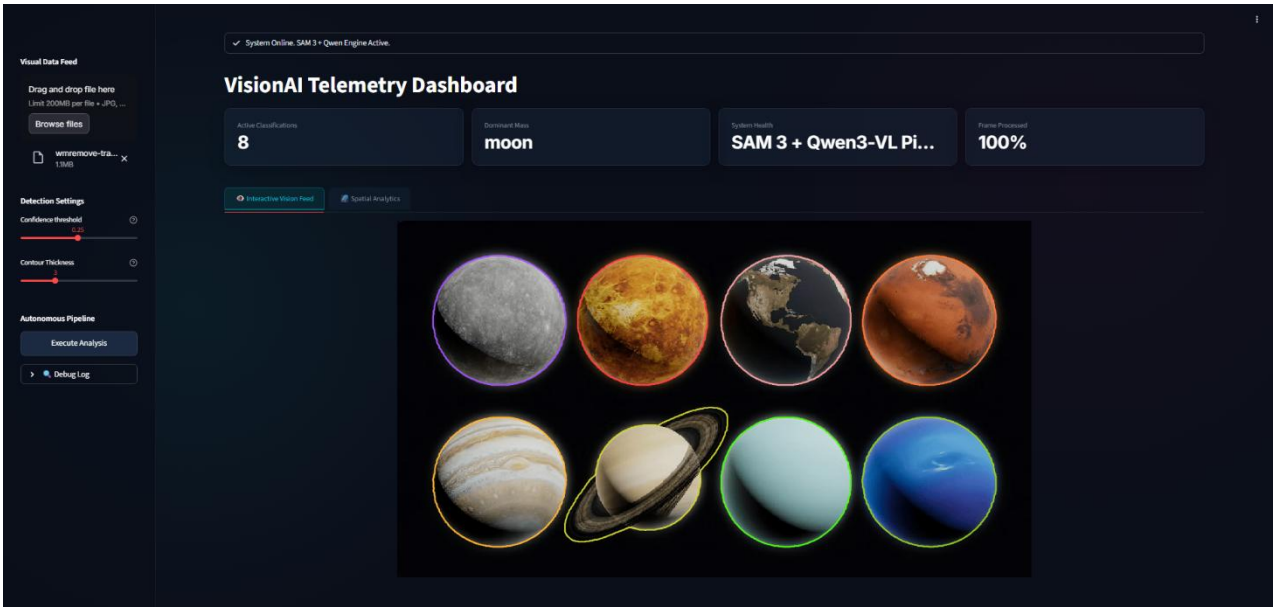
Результати дослідження

Тестування розробленого програмного забезпечення показало високу ефективність запропонованого підходу для розпізнавання об'єктів без попередньо заданих класів. Система автоматично генерує маски та розраховує ключові просторові метрики для кожного сегмента: площу у пікселях, відсоток від загальної площі (Volume %) та координати центрів мас.

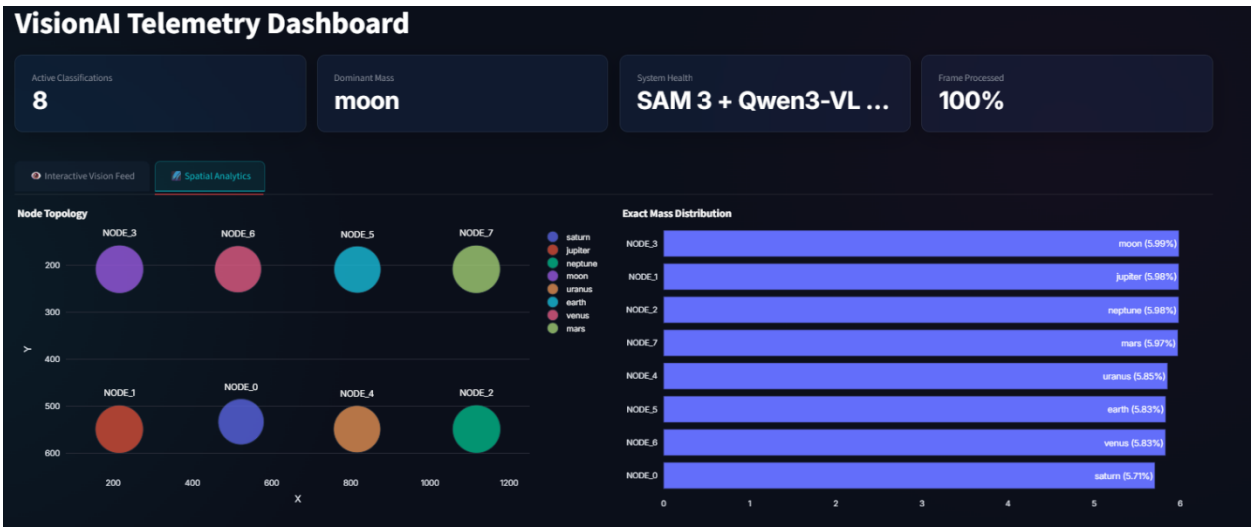
Оцінка результатів інтегрована у вкладку «Spatial Analytics», де за допомогою бібліотеки Plotly реалізовано два ключові аналітичні інструменти:

1. Топологія вузлів (Node Topology) - діаграма розсіювання, що відображає просторове розташування об'єктів на площині.
2. Розподіл маси (Mass Distribution) - гістограма, що ранжує ідентифіковані класи за площею, яку вони займають на зображенні.

Аналіз підтвердив, що використання Qwen3-VL дозволяє точно ідентифікувати семантичний контекст сцени, тоді як SAM 3 забезпечує безпрецедентну точність виділення складних границь об'єктів, відфільтровуючи незначні шумові контури (менше 0.2% від загальної площі). Зображення на рисунку 2 демонструють результати роботи програмного комплексу.



А1)



A2)

B1)

B2)

Рисунок 2 - A1) Сегментація астрономічних об'єктів, A2) Візуалізація аналізу зображення; B1) Семантична сегментація МРТ, B2) Візуалізація аналізу МРТ-скану

Висновки

Розроблена гібридна асиметрична архітектура, що поєднує можливості Qwen3-VL та SAM 3, дійсно демонструє високий рівень ефективності у вирішенні завдань семантичної сегментації. Об'єднання потужної мультимодальної мовної моделі з передовим сегментаційним фреймворком дозволяє компенсувати слабкі сторони кожного з підходів окремо, створюючи синергію для точного виділення об'єктів на зображенні.

Ключовою перевагою запропонованого рішення є використання підходу з відкритим словником, що принципово усуває обмеження традиційних моделей, які працюють лише з фіксованим набором класів. Це забезпечує унікальну гнучкість при аналізі невідомих або динамічних сцен. Додатково, впровадження інтегрованого аналітичного дашборда суттєво оптимізує процес отримання просторової телеметрії, перетворюючи складний технічний пайплайн на зручний інструмент для вилучення кількісних метрик.

Завдяки поєднанню адаптивності, точності та зручності інтерпретації результатів, запропонована система має значний потенціал для широкого спектру прикладних галузей. Зокрема, вона може бути ефективно інтегрована в системи автономної навігації та робототехніки для розуміння оточення, у медичну діагностику для аналізу знімків, а також у супутниковий моніторинг для автоматизованого аналізу великих територій. Таким чином, архітектура виступає універсальним фундаментом для створення інтелектуальних систем візуального аналізу наступного покоління.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Електронне джерело «Segment Anything Model 3 (SAM 3)» URL: https://github.com/facebookresearch/sam3
2. Електронне джерело «Qwen-VL: Versatile Vision-Language Model» URL: https://github.com/QwenLM/Qwen-VL
3. Електронне джерело «vLLM: Easy, fast, and cheap LLM serving» URL: https://docs.vllm.ai/
4. Електронне джерело «Streamlit: The fastest way to build and share data apps» URL: https://streamlit.io/
5. Електронне джерело «RAG у 2026: від PoC до production — повний гайд» URL: https://webscraft.org/blog/rag-u-2026-vid-poc-do-production-povniy-gayd

Кривошей Іван Максимович – студент кафедри автоматизації та інтелектуальних інформаційних технологій, Вінницький національний технічний університет, м. Вінниця, e-mail: ivankrivoshey2003@gmail.com

Софина Ольга Юріївна **–** к.т.н., доцент кафедри АІІТ, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: olsofina@gmail.com

Kryvoshei Ivan Maksymovych - Student of the Department of Automation and Intelligent Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: ivankrivoshey2003@gmail.com

Olga Sofina – PhD, Associate Professor of Automation and Intelligent Information Technologies, Faculty of Computer Systems and Automatics Vinnytsia National Technical University, Vinnytsia, e-mail: olsofina@gmail.com