УДК: 004.932 + 004.89 + 549.9
DOI: 10.25686/978-5-8158-2474-4-2025-595-606

Сегментация изображений геологических аншлифов с использованием дообученных визуально-языковых моделей

С. Д. Загайнов[1], Д. М. Коршунов[2], А. В. Хвостиков[1]
[1] Московский государственный университет имени М.В. Ломоносова, Москва, Россия
[2] Геологический институт РАН, Москва, Россия

Аннотация. В работе предложен метод семантической сегментации минералов на изображениях геологических аншлифов, основанный на текстовых описаниях. Стандартные подходы к сегментации с открытым словарём (Open-Vocabulary Segmentation), использующие модели типа CLIP, показывают низкую эффективность на узкоспециализированных данных из-за того, что обучающие выборки общего назначения, на которых они обучались, не содержат достаточного количества примеров из целевой предметной области. Для решения этой проблемы предложен двухэтапный подход: сначала модель CLIP дообучается на целевом домене с использованием набора пар «изображение-текст», собранных с ресурса Mindat.org. Затем дообученная модель интегрируется в качестве энкодера признаков в фреймворк Trident, не требующий дополнительного обучения. Эксперименты показывают, что предложенный подход значительно улучшает качество сегментации по сравнению с использованием базовой модели CLIP. Метрика F1 для классификации минералов на тестовых изображениях выросла с 0.16 до 43, а индекс Жаккара – с 0.10 до 0.29.

Ключевые слова: семантическая сегментация, сегментации с открытым словарём, геологические аншлифы, CLIP, глубокое обучение, компьютерное зрение.

Segmentation of geological polished section images using fine-tuned vision-language models

S. D. Zagaynov[1], D. M. Korshunov[2], A. V. Khvostikov[1]
[1] Lomonosov Moscow State University, Moscow, Russia
[2] Geological Institute, Russian Academy of Sciences, Moscow, Russia

Abstract. This paper proposes a method for semantic segmentation of minerals in geological thin section images based on text descriptions. Standard Open-Vocabulary Segmentation approaches using CLIP-like models show low performance on highly specialized data due to the lack of relevant knowledge in general-purpose training datasets. To address this issue, a two-stage approach is proposed: first, the CLIP model is fine-tuned on the target domain using a custom dataset of image-text pairs collected from Mindat.org. Second, the fine-tuned model is integrated as a feature encoder into the state-of-the-art Training-Free framework, Trident. Experiments demonstrate that the proposed approach significantly improves segmentation quality compared to using the baseline CLIP model. The F1-score for mineral classification on test images increased from 0.16 to 0.43, and the Jaccard index increased from 0.10 to 0.29.

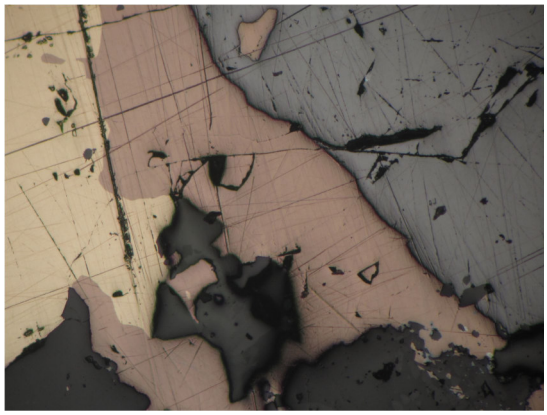
Keywords: semantic segmentation, open-vocabulary segmentation, geological thin sections, CLIP, deep learning, computer vision.

Введение

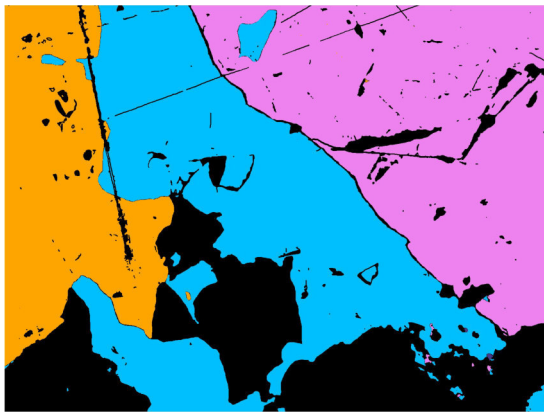
Семантическая сегментация – задача разделения изображения на семантически значимые области – является фундаментальной проблемой в компьютерном зрении с широким спектром применений. Цель состоит в том, чтобы для каждого пикселя изображения определить, к какому классу объектов он принадлежит. Однако традиционные подходы к семантической сегментации сталкиваются с существенными ограничениями. Во-первых, они оперируют в рамках закрытого набора предопределенных классов, что делает невозможным сегментацию новых, ранее не встречавшихся категорий. Во-вторых, создание точных попиксельных масок для обучения таких моделей представляет собой чрезвычайно трудоемкий и дорогостоящий процесс.

Эта проблема особенно актуальна для специализированных областей, таких как анализ геологических аншлифов, где важно иметь попиксельную карту для изображений аншлифов, в которой каждый цвет соответствует определенному минералу (рис. 1). Однако создание таких эталонных масок,

как на примере из набора LumenStone[1], требует глубоких экспертных знаний в области минералогии и является главным препятствием для использования стандартных методов обучения с учителем.



а) Исходное изображение аншлифа



б) Маска семантической сегментации

Рис. 1. Пример желаемого результата семантической сегментации для геологического изображения. Каждому цвету на маске *(б)* соответствует определенный минерал, присутствующий на исходном изображении *(а)*. Данные из набора LumenStone S1

В настоящее время уже предложено несколько сегментационных моделей, построенных на классических методах компьютерного зрения (ResUNet [1, 2]), однако такие модели способны лишь определять классы минералов, а задачи структурно-текстурного анализа остаются за пределами их возможностей. Хотя именно на выделении характерных «текстурных рисунков» строятся генетические интерпретации процессов рудообразования.

Для преодоления этих ограничений активно развивается направление семантической сегментации с открытым словарём (Open-Vocabulary Semantic Segmentation, OVS). Цель OVS – создание моделей, способных сегментировать на изображении объекты любых произвольных классов, заданных текстовым описанием, а не только тех, которые были заранее определены в обучающем наборе данных. Это достигается за счёт использования мощных предварительно обученных визуально-языковых моделей (Vision Language Models, VLMs), таких как CLIP [3], которые научились сопоставлять визуальные концепции с текстом на огромных массивах данных из интернета.

Несмотря на успехи, прямая адаптация моделей типа CLIP для задач сегментации сопряжена с трудностями. Их архитектура нацелена на классификацию всего изображения и порождает пространственно-инвариантные признаки, что мешает точной локализации объектов. Современные подходы, не требующие дополнительного обучения (Training-Free), пытаются обойти это ограничение. Одним из передовых решений является фреймворк Trident [4], который эффективно комбинирует возможности CLIP (для семантического понимания), DINO [5] (для получения пространственно-локализованных признаков) и Segment Anything Model (SAM) [6] (для генерации высококачественных масок). Такие методы позволяют достичь высокой гибкости и снизить затраты на подготовку данных.

Однако общим недостатком всех методов, основанных на CLIP, является их зависимость от знаний, заложенных в модель при предварительном обучении. Если целевые классы объектов очень специфичны и редко встречаются на фотографиях из интернета (например, геологические минералы, видимые под микроскопом), то качество сегментации будет низким. Модель может неверно интерпретировать текстовые запросы или не находить соответствующие визуальные признаки.

Данная работа посвящена решению проблемы сегментации минералов на изображениях геологических аншлифов. Актуальность этой задачи подтверждается активными исследованиями в данной области, в том числе разработкой специализированных нейросетевых методов, использующих оптические свойства минералов при съемке в разных режимах поляризации для повышения точности [7]. Однако фундаментальным ограничением таких подходов является их зависимость от обучения с

[1] https://imaging.cs.msu.ru/en/research/geology/lumenstone

учителем, которое требует наличия больших наборов данных с точной попиксельной разметкой. Создание подобных эталонных наборов, как, например, LumenStone [1], само по себе является отдельной трудоёмкой научной задачей. Эта зависимость от дорогостоящей ручной разметки ограничивает масштабируемость и адаптацию существующих методов к широкому спектру разнообразных и редких минералов, что и обуславливает актуальность разработки принципиально иных подходов, исследуемых в данной работе.

Основной вклад авторов данной статьи заключается в разработке и оценке метода, который адаптирует передовой OVS-подход к узкоспециализированной геологической предметной области. В данной работе предлагается двухэтапный подход.

1. Дообучение (fine-tuning) модели CLIP на целевом домене с использованием специализированного набора данных пар «изображение-текст», собранного из геологических источников.
2. Интеграция дообученной модели в качестве энкодера признаков в фреймворк Trident для выполнения семантической сегментации минералов.

Обзор существующих методов

1. *Семантическая сегментация с открытым словарем без дополнительного обучения*

Подходы, не требующие дополнительного обучения (Training-Free), представляют особый интерес, поскольку они позволяют избежать затрат на аннотирование данных и сохраняют обобщающие способности исходных моделей.

Trident [4] является одной из передовых работ в этой области. Фреймворк направлен на решение проблем пространственной инвариантности и ограниченного разрешения CLIP при работе с изображениями высокого разрешения. Trident предлагает парадигму «Splice-then-Segment»: сначала извлекаются признаки из частей изображения с помощью CLIP и DINO, а затем эти признаки объединяются в единую карту. Ключевым шагом является использование энкодера SAM для создания матрицы корреляции, которая применяется к этой карте, обеспечивая глобальную агрегацию информации. Это позволяет успешно сегментировать объекты даже на изображениях высокого разрешения. Кроме того, Trident предлагает стратегию уточнения грубых результатов сегментации, преобразуя их в промпты (точки, рамки, маски) для декодера SAM.

Другой Training-Free подход представлен в работе TextRegion [8]. Вместо работы с признаками на уровне отдельных пикселей или патчей, этот метод сначала использует мощную модель сегментации (например, SAM2 [9]) для разделения изображения на множество регионов-кандидатов. Затем для каждого региона признаки из визуально-языковой модели (например, CLIP) агрегируются (путем взвешенного суммирования) в единый «региональный токен». Этот токен напрямую сравнивается с эмбеддингом текстового запроса, превращая задачу сегментации в задачу классификации регионов. Однако применимость этого подхода к нашей задаче анализа геологических аншлифов ограничена, что проявляется в двух основных аспектах:

- зависимость от качества начальной сегментации. Эффективность метода полностью определяется тем, насколько хорошо SAM2 может выделить все потенциально значимые минералы. На сложных текстурах аншлифов, где границы между минералами могут быть нечеткими или сложными, SAM2 может допускать ошибки, которые невозможно будет исправить на последующих этапах;
- проблема отсутствия фона. Подход TextRegion использует специальные техники, чтобы отфильтровать признаки, относящиеся ко всему изображению в целом (глобальные признаки), и сосредоточиться на локальных. На изображениях аншлифов часто нет фона в привычном понимании: вся область снимка заполнена разными минералами. В таких условиях, во-первых, фильтрация глобальных признаков может работать некорректно. Во-вторых, усреднение признаков внутри маски, которая по ошибке захватила несколько разных минералов, приведет к потере уникальных текстурных характеристик каждого из них. В результате модель не сможет правильно классифицировать такой регион.

2. *Слабоконтролируемая семантическая сегментация с открытым словарем*

Методы этого типа стремятся снизить зависимость от полной попиксельной разметки, используя более доступные формы контроля, такие как пары изображение-текст или неразмеченные данные.

OVSegmentor [10] предлагает решение, обучаемое исключительно на парах «изображение-текст». В его основе лежит трансформерная архитектура, которая с помощью механизма slot-attention [11] группирует визуальные патчи в групповые токены, представляющие семантические области, и сопоставляет их с эмбеддингами текста. Для обучения используются прокси-задачи, такие как «завершение замаскированных сущностей», что заставляет модель устанавливать точное соответствие между визуальными группами и текстовыми понятиями, и «согласованность масок между изображениями»: модель учится генерировать схожие сегментационные маски для одной и той же сущности на разных изображениях.

CLIP-DINOiser [12] улучшает локализационные способности признаков из MaskCLIP [13], интегрируя в них информацию от модели DINO. Для этого обучаются легковесные свёрточные слои, которые учатся предсказывать «карты объектности» и улучшать пространственные характеристики признаков CLIP, используя DINO и метод FOUND [14] в качестве «учителей» на неразмеченных данных.

TCL [15] решает проблему несоответствия между обучением на уровне всего изображения и сегментацией на уровне регионов. Ключевым элементом TCL является обучаемый декодер, выполняющий функцию семантической локализации (text grounding). Он генерирует маску, указывающую на области изображения, которые семантически связаны с данным текстом. Визуальные эмбеддинги из этих областей напрямую сопоставляются с текстом с помощью контрастивной функции потерь. Однако данный метод плохо подходит для задачи сегментации минералов, так как его функция потерь предполагает, что на изображении присутствует только один целевой объект, и наказывает за наличие других объектов, которые могут быть целевыми в других примерах батча (пакета).

Общий вывод из анализа существующих решений заключается в том, что, несмотря на прогресс, их эффективность в узкоспециализированных областях остаётся ограниченной из-за отсутствия этих знаний у базовой модели CLIP. Наша работа нацелена на устранение именно этого недостатка.

Методология

Предлагаемый в данной работе метод состоит из трёх основных этапов: сбор и подготовка данных для геологического домена, дообучение модели CLIP на этих данных и интеграция адаптированной модели в фреймворк Trident для выполнения сегментации.

1. *Сбор и подготовка набора данных*

Для адаптации модели к предметной области был сформирован специализированный набор данных. Источником послужила онлайн-база по минералогии Mindat.org[1]. Основное внимание было уделено изображениям полированных срезов (аншлифов) и тонких срезов (шлифов) из категорий «Polished Section (Polarized Light)» и «Polished Section (Slice or Surface)», каждый такой снимок сопровождается списком присутствующих минералов и описанием.

Процесс сбора и обработки данных включал следующие шаги.

1. Сбор «сырых» данных: с помощью автоматизированного парсинга были извлечены изображения, их оригинальные текстовые описания и списки присутствующих минералов.

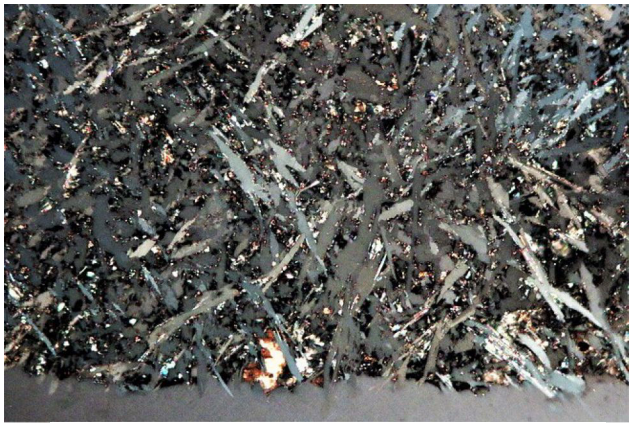
2. Извлечение названий минералов: с помощью регулярных выражений из текстовой информации были извлечены наименования минералов.

3. Фильтрация классов: был проведён частотный анализ, по результатам которого отобраны топ-20 наиболее часто встречающихся минералов (Arsenopyrite, Bismuth, Bornite, Calcite, Chalcedony, Chalcocite, Chalcopyrite, Covellite, Dolomite, Galena, Goethite, Hematite, Magnetite, Nickeline, Pyrite, Pyrrhotite, Quartz, Rammelsbergite, Sphalerite, Tetrahedrite Subgroup). Это было сделано для обеспечения достаточного количества примеров для каждого класса и упрощения задачи на начальном этапе исследования.

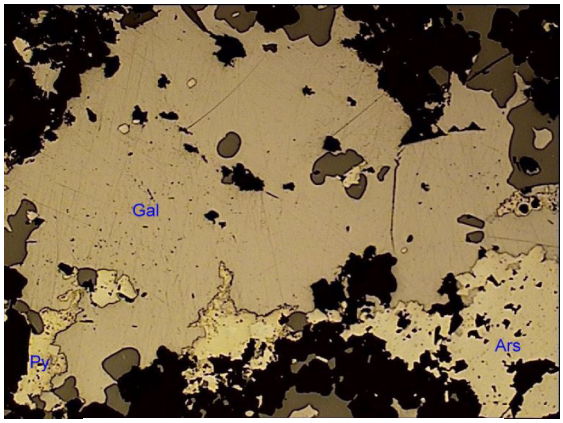
[1] https://www.mindat.org/

4. Формирование текстовых описаний: для каждого изображения, содержащего отобранные минералы, было сформировано стандартизированное описание: *"Minerals on this photo: <минерал_1, минерал_2, ...>. Description: <оригинальное текстовое описание>"*. Такой формат предоставляет модели как явную информацию о целевых классах, так и дополнительный геологический контекст.

В результате был сформирован набор данных, состоящий из 9884 пар «изображение-текст», который далее использовался для дообучения модели CLIP. Примеры таких пар показаны на рисунке 2.



***а*)** Minerals on this photo: Galena. Description: Felted acicular zinkenite crystals (greenish- to blueish- to brownish-grey anisotropy with some specs of red internal reflections) in association with galena (grey, bottom). Vertically reflected plane polarized light digital image in air, with crossed polarizers, width 0.5 mm



***б*)** Minerals on this photo: Galena, Pyrite, Arsenopyrite. Description: POlished section with galena (Gal) and some pyrite (Py) and arsenopyrite (Ars)

Рис. 2. Примеры пар «изображение-текст» из собранного набора данных (текстовые описания изображений на английском языке)

2. *Архитектура и дообучение модели CLIP*

Модель CLIP [3] состоит из двух энкодеров: визуального (Image Encoder) и текстового (Text Encoder), которые преобразуют изображения и текст в эмбеддинги в общем векторном пространстве. В данной работе в качестве визуального энкодера использовалась архитектура Vision Transformer (ViT-B/16) [16], а в качестве текстового – трансформер на основе архитектуры GPT-2 [17].

Обучение модели основано на контрастивном подходе, целью которого является максимизация сходства между эмбеддингами соответствующих пар «изображение-текст» и минимизация для несоответствующих. Для пакета из N пар вычисляется матрица косинусных сходств S размером N×N. Затем вычисляется симметричная контрастивная функция потерь:

$$L_{\text{img}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ij}/\tau)},$$

$$L_{\text{text}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N}\exp(S_{ji}/\tau)},$$

$$L_{\text{CLIP}} = \frac{L_{\text{img}} + L_{\text{text}}}{2},$$

где τ – обучаемый параметр.

Для эффективной адаптации модели с меньшими вычислительными затратами и для снижения риска катастрофического забывания была применена техника параметро-эффективной настройки (Parameter-Efficient Fine-Tuning, PEFT), а именно метод LoRA (Low-Rank Adaptation) [18]. Ключевая идея LoRA заключается в модификации выходного вектора h каждого адаптируемого слоя путем добавления к нему низкоранговой поправки. Вместо прямого дообучения исходной матрицы весов W_0 её оставляют замороженной, а выход слоя вычисляется по формуле

$$h = W_0 x + \frac{\alpha}{r} BAx.$$

В данной формуле x – это входной вектор, а W_0x – результат работы исходного, предварительно обученного слоя. Адаптация достигается за счёт второго члена, где обучаемыми являются только две новые, низкоранговые матрицы $A \in \mathbb{R}^{r\times k}$ и $B \in \mathbb{R}^{d\times r}$. Ключевой гиперпараметр, ранг адаптации r (где $r << C \min(d, k)$), определяет внутреннюю размерность этих матриц, то есть контролирует количество обучаемых параметров и сложность аппроксимации. В работе было выбрано стандартное значение $r = 32$. Второй гиперпараметр, α, является коэффициентом масштабирования, который регулирует итоговую величину вклада от адаптера, позволяя управлять балансом между исходными знаниями модели и новыми. Использовалось стандартное значение $\alpha = 64$.

LoRA-адаптеры с указанными гиперпараметрами были добавлены к линейным слоям в блоках внимания и полносвязных слоях (MLP) обоих энкодеров: визуального и текстового. Дообучение проводилось 5 эпох с размером пакета 32, начальной скоростью обучения 1e-6, оптимизатором AdamW [19] и планировщиком OneCycleLR [20].

3. *Интеграция в фреймворк Trident*

Дообученная модель CLIP была интегрирована в фреймворк Trident, заменив собой базовую модель. Trident реализует парадигму Splice-then-Segment, общая архитектура которой показана на рисунке 3. Метод состоит из четырёх основных этапов, которые были адаптированы для рассматриваемой задачи.

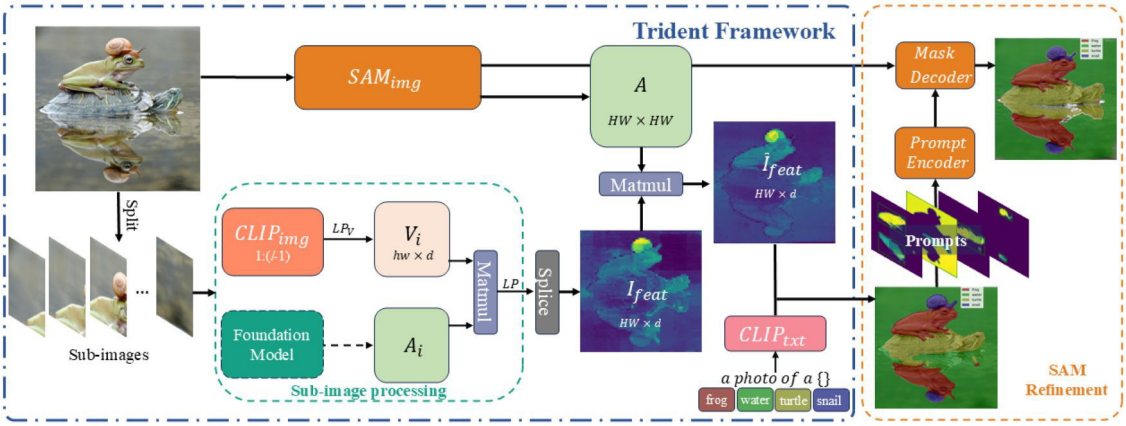


Рис. 3. Общая архитектура фреймворка Trident. Дообученная модель CLIP используется на этапе извлечения семантических признаков, которые затем обрабатываются и уточняются с помощью DINO и SAM (иллюстрация из статьи [4])

1. **Обработка субизображений**. Исходное изображение высокого разрешения I^{src} делится на множество n перекрывающихся субизображений (патчей) с помощью скользящего окна. Размер патчей соответствует входному разрешению энкодера CLIP (224×224 пикселя), а перекрытие обеспечивает сохранение контекста между соседними областями. Для каждого патча извлекаются два типа признаков: семантические признаки V_i с помощью энкодера дообученной модели CLIP и пространственно-локализованные признаки A_i; от модели DINO. Эти признаки объединяются, чтобы дополнить высокоуровневые семантические признаки пространственной детализацией, и затем соединяются в одну общую карту признаков I^{feat}.

2. **Глобальная агрегация**. Энкодер SAM обрабатывает полное исходное изображение I^{src} для генерации корреляционной матрицы A, которая улавливает семантические отношения между пикселями. Эта матрица A затем умножается на общую карту признаков I^{feat}, агрегируя информацию в глобальном масштабе: $\tilde{I}^{feat} = A \cdot I^{feat}$, где $\cdot$ – матричное умножение. Это позволяет эффективно расширить рецептивное поле и учесть контекст всего изображения.

3. **Сегментация**. Агрегированные признаки $\tilde{I}^{feat}$ сравниваются с текстовыми эмбеддингами T_k^{emd} для каждого целевого класса k (названия минерала), которые также генерируются текстовым энкодером (дообученной моделью CLIP). Начальная карта сегментации S получается путём присвоения каждому пикселю метки класса с максимальным косинусным сходством:

$$S(x,y) = \underset{k}{\operatorname{argmax}} \cos(\widetilde{I}^{feat}(x,y), T_k^{emd}).$$

4. **Уточнение с помощью SAM**. Полученная карта сегментации S может быть грубой и неточной по краям. Для её уточнения Trident использует декодер и промпт-энкодер модели SAM. Поскольку SAM наиболее эффективен при работе с точечными и рамочными промптами, Trident преобразует начальные маски для каждого класса в комбинацию из трёх типов промптов.

1) ***Точечные промпты***: точки внутри найденной области, где уверенность модели в предсказании класса максимальна. Это помогает SAM сфокусироваться на наиболее репрезентативных частях объекта.

2) ***Рамочные промпты***: минимальный ограничивающий прямоугольник (bounding box), который полностью охватывает найденную область (задаёт общие границы объекта).

3) ***Масочные промпты***: грубая бинарная маска

$$B_k(x,y) = \begin{cases} 1, & \text{если } S(x,y) = k \\ 0, & \text{иначе} \end{cases},$$

которая указывает на примерное расположение и форму объекта.

Одновременная подача этих промптов в SAM позволяет сгенерировать финальные, значительно более точные и качественные, маски для каждого объекта.

Эксперименты и результаты

1. *Детали реализации*

Все эксперименты по дообучению модели проводились на одном графическом ускорителе NVIDIA A6000. Процесс дообучения на собранном наборе данных длился 10 эпох и занял около 3 часов. Оценка производительности и визуализация результатов выполнялись на той же аппаратной конфигурации.

2. *Настройка экспериментов*

Эксперименты проводились для сравнения двух конфигураций.

1) ***Базовый Trident***: использует оригинальную модель CLIP (ViT-B/16), обученную на LAION-400M.

2) ***Предложенный метод (Trident + Fine-tuned CLIP):*** использует модель CLIP, дообученную на наборе данных, полученном на основе Mindat.org.

Оценка проводилась на отложенной тестовой выборке из собранного набора данных. Для каждого изображения в качестве текстовых запросов подавались названия минералов, которые, согласно разметке, присутствуют на нём. Таким образом, задача семантической сегментации минералов при таком упрощении сводится к задаче классификации.

3. *Метрики оценки*

Поскольку получение точных попиксельных масок для тестового набора затруднительно, то задача оценивалась как мультиклассовая классификация на уровне всего изображения. Для каждого изображения получался набор предсказанных классов (минералов, для которых были сгенерированы непустые маски) и сравнивался его с истинным набором классов. Использовались следующие метрики:

1) ***Jaccard (индекс Жаккара):*** $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, где A – набор предсказанных, B – набор истинных классов;

2) ***Recall (полнота):*** $R = \frac{TP}{TP+FN}$, доля найденных истинных классов;

3) ***Precision (точность):*** $P = \frac{TP}{TP+FP}$, доля верных среди предсказанных классов;

4) ***F1-score***: $F1 = 2 \cdot \frac{P \cdot R}{P+R}$, гармоническое среднее точности и полноты.

4. *Количественные результаты*

Результаты количественной оценки эффективности предложенного метода на отложенной тестовой выборке из собранного нами набора данных с Mindat.org представлены в таблицах 1 и 2.

Таблица 1. **Метрики классификации для базового подхода Trident**

Название метрики	*Значение*
Jaccard	0.1006
Recall	0.3106
Precision	0.1278
F1-score	0.1645

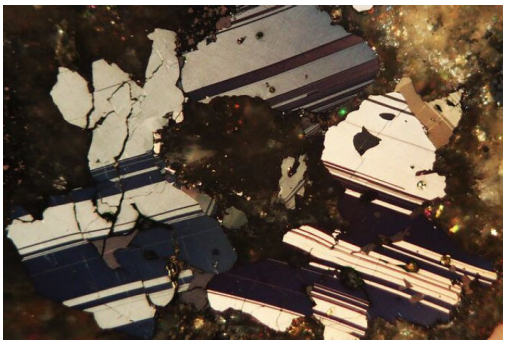
Таблица 2. **Метрики классификации для предложенного метода**

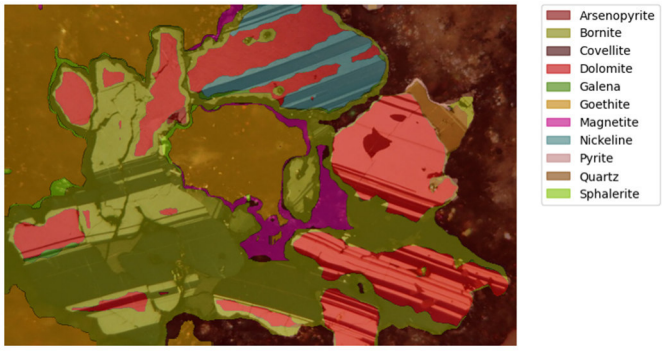
Название метрики	*Значение*
Jaccard	0.2929
Recall	0.5795
Precision	0.4405
F1-score	0.4326

Таблицы показывают, что дообучение модели CLIP на целевом домене привело к значительному улучшению всех метрик. F1-score увеличился (с 0.1645 до 0.4326), что указывает на существенное повышение способности модели правильно идентифицировать минералы. Увеличение Recall говорит о том, что модель стала реже пропускать присутствующие минералы, а рост Precision – о снижении количества ложных срабатываний.

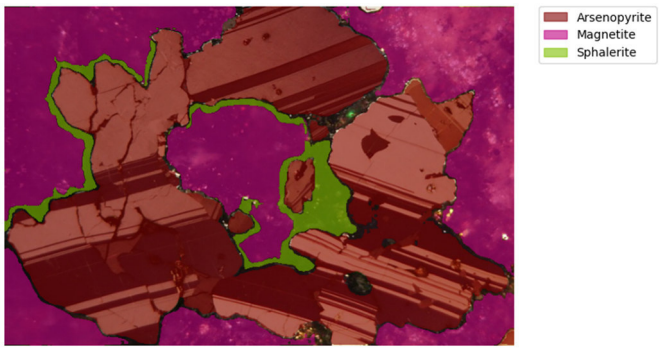
5. *Визуальные результаты*

Визуализация результатов (рис. 4-7) подтверждает полученные количественные оценки. На них продемонстрировано, как дообучение улучшает качество сегментации.



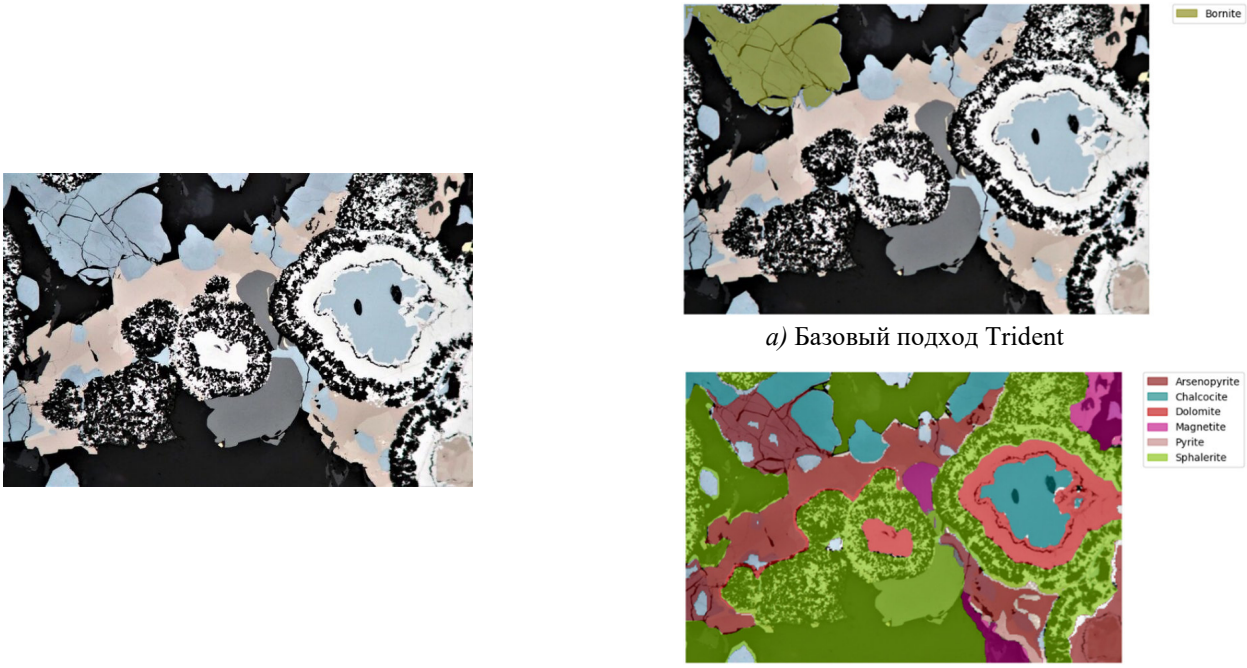


а) Базовый подход Trident



б) Предложенный метод

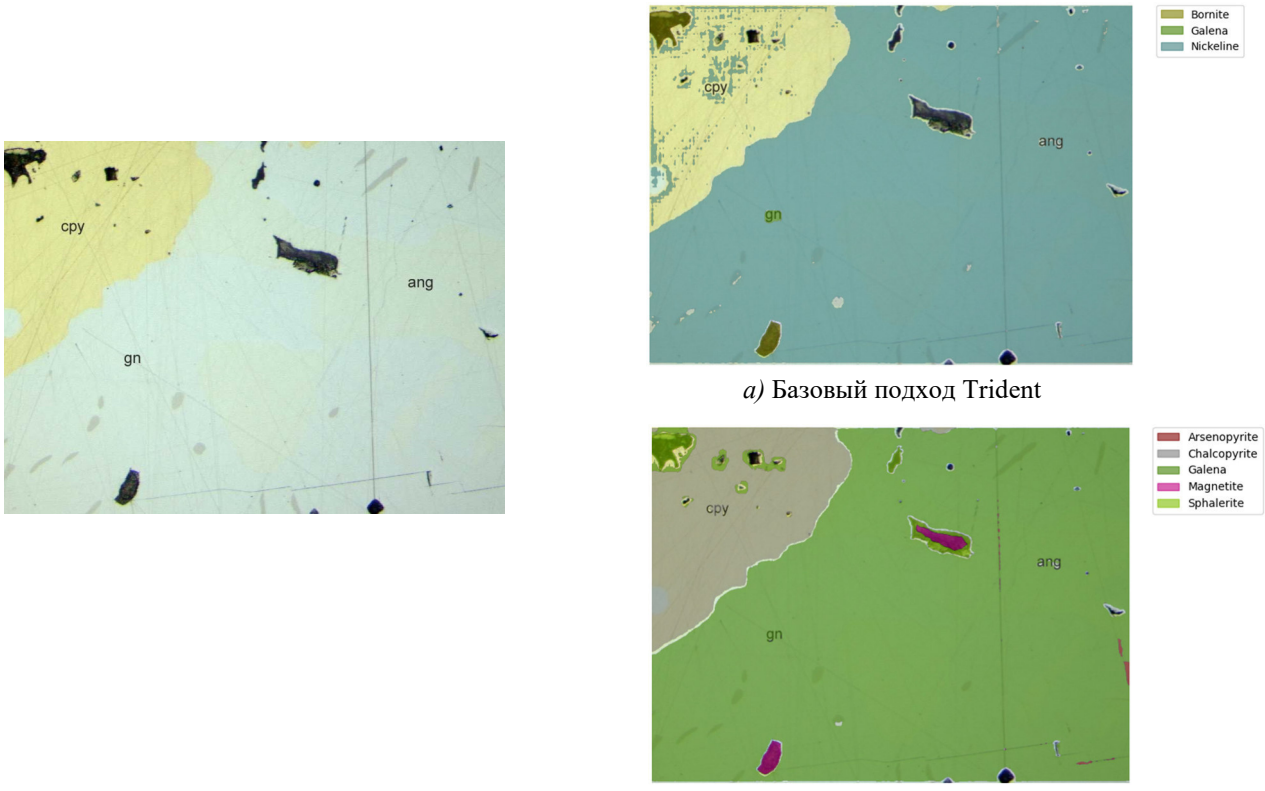
Рис. 4. Минералы, присутствующие на изображении: Sylvanite, Pyrite, Galena, Quartz. *Слева* – исходное изображение. *Справа* – изображения с наложенной маской сегментации



а) Базовый подход Trident

б) Предложенный метод

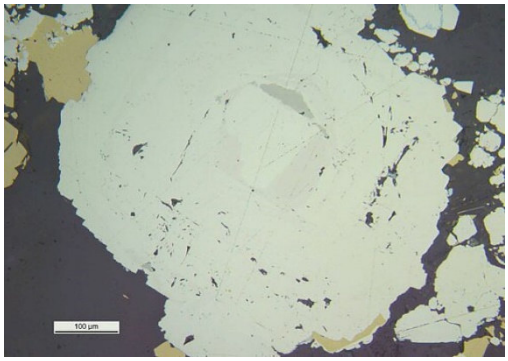
Рис. 5. Минералы, присутствующие на изображении: Chalcopyrite, Sphalerite. *Слева* – исходное изображение. *Справа* – изображения с наложенной маской сегментации

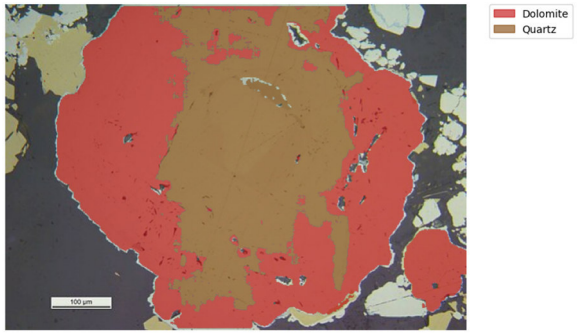


а) Базовый подход Trident

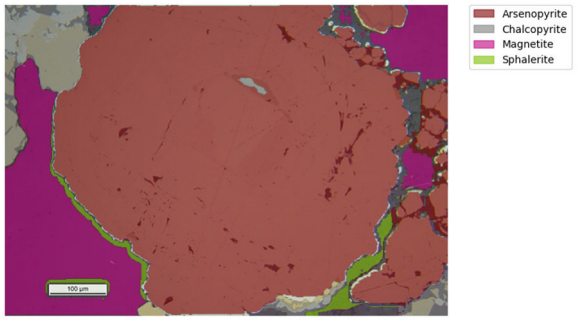
б) Предложенный метод

Рис. 6. Минералы, присутствующие на изображении: Chalcopyrite, Galena. *Слева* – исходное изображение. *Справа* – изображения с наложенной маской сегментации





а) Базовый подход Trident



б) Предложенный метод

Рис. 7. Минералы, присутствующие на изображении: Pyrite, Chalcopyrite. *Слева* – исходное изображение. *Справа* – изображения с наложенной маской сегментации

Анализ визуальных результатов (см. рис. 4-7) позволяет качественно оценить преимущества предложенного подхода. На примерах видно, что дообученная модель демонстрирует более глубокое понимание геологического контекста, текстур и структур.

На рисунке 4 представлено изображение со сложной структурой, базовый метод *(а)* генерирует сильно фрагментированную маску с большим количеством ложных классов. В то же время предложенный метод *(б)* лучше справляется с задачей, формируя семантически верные маски, но также ошибается в определении классов.

На рисунке 5 базовый подход Trident *(а)* не может распознать сложную структуру, в то время как предложенный метод *(б)* сегментирует различные минералы, что говорит о его способности определять не только цвет, но и геологические паттерны.

Особенно ярко преимущество предложенного метода *(а)* видно на рисунке 6. Базовый метод допускает ошибки в классификации и генерирует шумные маски, тогда как предложенный метод *(б)* правильно определяет основные минералы, присутствующие на изображении.

На рисунке 7 базовый метод *(а)* ошибочно разделяет крупный фрагмент минерала на несколько разных классов, нарушая его целостность. Предложенный метод *(б)*, напротив, корректно идентифицирует весь фрагмент как единый объект.

Таким образом, визуальные результаты показывают, что дообучение CLIP позволяет модели не просто лучше классифицировать минералы, но и генерировать более точные и геологически осмысленные маски сегментации, сохраняя целостность объектов и распознавая сложные структуры.

Заключение

В данной работе был предложен и исследован метод семантической сегментации для узкоспециализированной области – анализа изображений геологических аншлифов. Ключевой идеей метода является адаптация мощной визуально-языковой модели CLIP к целевому домену путём дообучения на специально собранном наборе данных, с последующей её интеграцией в передовой OVS-фреймворк Trident.

Проведённые эксперименты убедительно доказывают применимость предложенного метода визуально-языковой модели сегментации для целей анализа изображений аншлифов. Дообучение CLIP на доменных данных значительно повышает качество сегментации по сравнению с использованием базовой, универсальной модели. Это подтверждается как существенным ростом количественных метрик (F1-score вырос с 0.16 до 0.43), так и улучшением визуальных результатов сегментации. Предложенный подход позволяет эффективно адаптировать существующие OVS-решения для задач в специализированных областях, где сбор больших наборов данных с попиксельной разметкой затруднён или невозможен.

При использовании данной сегментационной модели совместно с уже разработанными традиционными методами компьютерного зрения (например, ResUNet[1]), обученными на больших наборах данных аннотированных изображений аншлифов, в перспективе получится создать полноценный механизм автоматического описания руд под микроскопом, который будет способен не только определять минералы и статистически обрабатывать их распределения (содержания и группы по размерам), но и проводить полноценный структурно-текстурный анализ.

Предлагаемый метод был программно реализован на языке Python с использованием библиотек PyTorch, NumPy, Pillow, Skimage, а также открытых реализаций OpenCLIP и Trident.

Дальнейшее развитие

Несмотря на положительные результаты, работа на текущий момент имеет некоторые ограничения. Во-первых, набор данных ограничен 20 наиболее частыми минералами. Во-вторых, оценка проводилась на уровне классификации изображений, а не попиксельного совпадения масок, из-за отсутствия точной разметки.

В качестве дальнейшего развития работы можно выделить следующие направления:

- расширение набора данных для включения более редких минералов;
- создание небольшого валидационного набора с точной попиксельной разметкой для оценки качества сегментационных масок с помощью метрики mIoU;
- исследование влияния различных архитектур VLM на итоговое качество сегментации минералов.

Источник финансирования

Исследование выполнено за счёт гранта Российского научного фонда (проект № 24-21-00061).

Список литературы

1. Automatic identification of minerals in images of polished sections / AV Khvostikov, DM Korshunov, AS Krylov, MA Boguslavskiy // *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences.* 2021. Vol. 44. Pp. 113-118.

2. An improved mineral image recognition method based on deep learning / Huaming Tang, Hongming Wang, Ling Wang, et al. // *Jom.* 2023. Vol. 75, no. 7. Pp. 2590-2602.

3. Learning transferable visual models from natural language supervision / Alec Radford, Jong Wook Kim, Chris Hallacy, et al. // International conference on machine learning / PmLR. 2021. Pp. 8748-8763.

4. *Shi Yuheng, Dong Minjing, Xu Chang.* Harnessing Vision Foundation Models for High-Performance, Training-Free Open Vocabulary Segmentation // *arXiv preprint arXiv:2411.09219.* 2024.

5. Emerging properties in self-supervised vision transformers / Mathilde Caron, Hugo Touvron, Ishan Misra, et al. // Proceedings of the IEEE/CVF international conference on computer vision. 2021. Pp. 9650-9660.

6. Segment anything / Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. // Proceedings of the IEEE/CVF international conference on computer vision. 2023. Pp. 4015-4026.

7. Registration and segmentation of PPL and XPL images of geological polished sections containing anisotropic minerals / DI Razzhivina, DM Korshunov, MA Boguslavsky, et al. // *Computational Mathematics and Modeling.* 2023. Vol. 34, no. 1. Pp. 16-26.

8. TextRegion: Text-Aligned Region Tokens from Frozen Image-Text Models / Yao Xiao, Qiqian Fu, Heyi Tao, et al. // *arXiv preprint arXiv:2505.23769.* 2025.

9. Sam 2: Segment anything in images and videos / Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. // *arXiv preprint arXiv:2408.00714.* 2024.

10. Learning open-vocabulary semantic segmentation models from natural language supervision / Jilan Xu, Junlin Hou, Yuejie Zhang, et al. // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023. Pp. 2935-2944.

11. *Object-Centric Learning with Slot Attention* / Locatello Francesco, Weissenborn Dirk, Unterthiner Thomas, et al. *2020.*

12. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free / Monika Wysoczanska, Michael Ramamonjisoa, Tomasz Trzcihski, Oriane Simeoni // Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. Pp. 1403-1413.

13. Learning to prompt for vision-language models / Kaiyang Zhou, Jingkang Yang, Chen Change Loy, Ziwei Liu // *International Journal of Computer Vision.* 2022. Vol. 130, no. 9. Pp. 2337-2348.

14. Unsupervised object localization: Observing the background to discover objects / Oriane Simeoni, Chloe Sekkat, Gilles Puy, et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. Pp. 3176-3186.

15. *Cha Junbum, Mun Jonghwan, Roh Byungseok.* Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. Pp. 11165-11174.

16. An image is worth 16x16 words: Transformers for image recognition at scale / Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. // *arXiv preprint arXiv:2010.11929.* 2020.

17. Language models are unsupervised multitask learners / Alec Radford, Jeffrey Wu, Rewon Child, et al. // *OpenAI blog.* 2019. Vol. 1, no. 8. P. 9.

18. Lora: Low-rank adaptation of large language models / Edward J Hu, Yelong Shen, Phillip Wallis, et al. // *ICLR.* 2022. Vol. 1, no. 2. P. 3.

19. Loshchilov Ilya, Hutter Frank. *Decoupled weight decay regularization //* arXiv preprint arXiv:1711.05101. *2017.*

20. *Smith Leslie N, Topin Nicholay.* Super-convergence: Very fast training of neural networks using large learning rates // Artificial intelligence and machine learning for multi-domain operations applications / SPIE. Vol. 11006. 2019. Pp. 369-386.