Содержание

1 Введение **4**

2 Обзорно-постановочный раздел работы **5**

2.1 Распознавание физических свойств объектов по их материалам 5

2.2 Датасеты с разметкой материалов объектов на изображениях 5

2.3 Методы распознавания материалов объектов на изображениях 7

2.4 Формальная постановка задачи 8

3 Метод **10**

3.1 Определение материалов на основе статистических признаков признаков 10

3.1.1 Статистические признаки изображений 11

3.1.2 Оценка важности признаков 14

3.1.3 Метрики качества классификации с несколькими метками . . . 15

3.2 Определение материалов с применением семантической сегментации . . 17

3.2.1 Модели семантической сегментации 17

3.2.2 Оценка качества моделей 18

3.3 Определение материалов с помощью визуально-языковых моделей . . . 19

3.4 Контрастивное обучение мультимодальной архитектуры 20

3.4.1 Общая функция потерь 24

3.4.2 Оценка качества работы модели 24

3.5 Архитектура модели 28

4 Разработка датасета MatLVIS **30**

4.1 Общее описание подхода к подготовке набора данных 31

4.2 Разметка с применением VLM 31

4.3 Валидация разметки с применением LLM 33

4.4 Доразметка датасета 35

4.5 Выделение категорий материалов из описаний объектов 35

4.6 Результаты разметки датасета 38

5 Вычислительные эксперименты **44**

5.1 Обучение классического алгоритма на датасете MatLVIS 44

5.2 Обучение моделей семантической сегментации на датасете DMS 45

5.3 Алгоритм определения материала объекта с помощью визуально-языковых моделей 47
5.4 Контрастивное обучение мультимодальной архитектуры на датасете MatLVIS 48
5.4.1 Выбор предобученной модели 48
5.4.2 Эксперименты с архитектурой модели 49
5.4.3 Эксперименты с функцией потерь 52
5.5 Сравнение различных методов распознавания материалов на датасете MatLVIS 56

6 Заключение **60**

Аннотация

Распознавание физических свойств объектов по изображениям является важной задачей компьютерного зрения и робототехники. Многие характеристики, такие как твердость, плотность и шероховатость, могут быть оценены на основе знания материалов объектов, однако существующие датасеты имеют ограниченный словарь материалов и недостаточное разнообразие естественных описаний.

В работе представлен комплексный подход к распознаванию материалов объектов в трехмерных сценах по RGB-изображениям. Ключевым результатом является создание датасета MatLVIS — расширения датасета LVIS, включающего корпус естественных описаний объектов с явным указанием материалов. Датасет содержит более 1.2 млн описаний и около 1600 уникальных категорий материалов, что значительно превышает возможности существующих датасетов.

Проведено сравнительное исследование различных подходов: статистических признаков, семантической сегментации, визуально-языковых моделей и контрастивного обучения. Предложена модифицированная мультимодальная архитектура на основе CLIP, объединяющая функции потерь InfoNCE и SupCon для учета сходства материалов и семантических классов. Архитектура дополнена модулем контекстной информации для улучшения эмбеддингов изображений.

Результаты работы имеют практическое значение для задач робототехники, требующих физического взаимодействия с объектами, а созданный датасет может служить основой для дальнейших исследований в области мультимодального машинного обучения и распознавания свойств материалов.

1 Введение

Распознавание физических свойств объектов полезно в любых задачах, где одним из этапов является физическое взаимодействие с объектами, например в задачах манипуляции. Многие физические свойства, такие как плотность, твердость, хрупкость, шероховатость и прочие, можно оценить, зная материал, из которого сделан объект. Существует не так уж много наборов данных, в которых разметка содержит информацию о материалах – и ещё меньше данных, в которых есть информация о физических характеристиках. Большая часть таких датасетов узкоспециализирована, что усложняет задачу обучения модели с достаточно общими знаниями о материалах или физических свойствах. Существующие наборы данных с информацией о физических свойствах сфокусированы на материалах, а не на объектах. В отличие от датасетов, применяющиеся в обозначенных задачах нейросетевые архитектуры, как правило, стандартные, если не считать мультимодальные решения. Такие модели обычно обрабатывают различные оптические модальности (RGB, инфракрасные снимки, поляризация) и объединяют выделенные признаки.

Для задачи определения материалов без специализированного домена хорошо подходят визуально-языковые модели: как правило, их языковая часть содержит достаточно знаний о мире для оценки качеств объекта, а визуальная часть способна выделить категорию объекта и дополнительную контекстную информацию. Главный недостаток визуально-языковых моделей – это их скорость работы. В домене задач манипуляции важны легкие и быстрые решения.

В данной работе рассмотрены различные методы распознавания материалов объекта по его изображению: подход на основе выделения классических признаков, семантическая сегментация, инференс визуально-языковых моделей. Предложена мультимодальная архитектура на основе CLIP, которая обучена строить информативные векторные представления изображений объектов. Предложен алгоритм применения такой модели для определения материалов объектов. Для обучения такой архитектуры разработан датасет MatLVIS, который дополняет датасет LVIS корпусом описаний, в которых в явном виде назван материал объекта. Предложен алгоритм автоматической разметки датасета с применением визуально-языковых и больших языковых моделей.

2 Обзорно-постановочный раздел работы

В данном разделе рассмотрены существующие наборы данных (2.1, 2.2) и методы (2.3) по распознаванию физических свойств объектов (в частности, материалов) по изображению. В подразделе 2.4 дана формальная постановка задачи распознавания материалов объектов по их изображениям.

2.1 Распознавание физических свойств объектов по их материалам

В работе [12] авторы предлагают использовать визуальные и тактильные данные для распознавания физических бинарных атрибутов объектов: cool, hard, metallic, nice, porous, rough, smooth, soft, solid, textured и других. Для этого разработан датасет PHAC-2, в котором представлены 530 различных объектов.

В датасете Describing Textures in the Wild (DTD) [7] представлены 5640 фотографий текстур, классифицированных по 47 "описательным" классам (например, striped, fibrous, grid, perforated, pleated).

Существуют и другие источники, в которых собраны данные по физическим свойствам материалов: например, открытая база данных Materials Project [5] и различные специализированные справочники.

На данный момент не существует наборов данных, в которых широкий набор количественных физических признаков был бы ассоциирован с объектом, однако, как будет показано в разделе 2.2, существует значительное количество датасетов с разметкой по материалам объектов. Данная работа также сфокусирована на распознавании материалов в предположении, что по материалу объекта с помощью дополнительных источников получится оценить его физические свойства.

2.2 Датасеты с разметкой материалов объектов на изображениях

Существует достаточно много датасетов для задачи распознавания и сегментации материалов.

Один из первых таких датасетов – UIUC Texture Database [4], который появился в 2005 году — классический набор из 1000 изображений и 25 категорий материалов,

по 40 изображений на класс. Изображения были сделаны в контролируемых лабораторных условиях с равномерным освещением и фиксированным ракурсом.

Похожий, но более поздний и разнообразный датасет – Flickr Material Database (FMD) [8]. Он содержит всего 1000 изображений и 10 категорий материалов. Изображения, как правило, содержат текстуру или объект крупным планом, в разметке также присутствуют бинарные маски.

Другой достаточно ранний датасет – OpenSurfaces [6], он содержит свыше 20 000 интерьерных фотографий с 100 000 сегментов по примерно 36 категориям.

Один из наиболее значимых датасетов – Materials in context (MINC) [10], содержащий около 437 000 изображений с 3 млн аннотаций точек по 23 категориям материалов в естественных сценах. В датасете есть бинарные маски объектов, однако их всего 7062.

Датасет Ground Terrain in Outdoor Scenes (GTOS) [14] сфокусирован на наземных поверхностях (>30 000 изображений, 31 класс) для автономной навигации.

Local Materials Database (LMD) [23] включает около 5000 фотоснимков локальных областей поверхностей для 19 категорий материалов. Каждое изображение снабжено метками местоположения масок материала и дополнительными статистическими признаками (градиенты, фильтры Габора).

Датасет Apple DMS [37] содержит разметку семантическими масками для 46 классов на 44 560 RGB-изображениях. Авторы работы также предлагают сверточную модель, которая достигает примерно 42% mIoU.

Существуют и синтетические наборы данных по материалам. Large Shape And Texture (LAS&T) [48] – это обширный синтетический датасет, включающий более 650000 изображений 2D/3D объектов в различных материалах. Каждая запись содержит фотореалистичные рендеры вместе с физическими свойствами материала.

MatPredict [47] — это датасет для изучения свойств материалов разнообразных объектов внутренней обстановки. Он включает 18 категорий объектов и в среднем по 14 рендеров на объект: каждый рендер снабжен картами нормалей, альбедо, глубины и параметров BRDF.

В таблице 1 приведены характеристики указанных в этом и предыдущем разделах датасетов. В дальнейшем в разделе 3.2 будем использовать датасет DMS, как наиболее объемный и содержащий большое число категорий.

Таблица 1: Существующие датасеты с разметкой по категориям материалов.

Датасет	Год	Кол-во изображений	Кол-во категорий	Тип разметки
UIUC Texture [4]	2005	1000	25	метки на уровне изображения
FMD [8]	2010	1000	10	метки на уровне изображения
OpenSurfaces [6]	2013	25357	36	бинарные маски
DTD [7]	2014	5640	47	ключевой атрибут + совместные атрибуты
MINC [10]	2015	436729	23	точечные аннотации + бинарные маски
PHAC-2 [12]	2016	5300	24	бинарные метки на уровне изображени
GTOS [14]	2018	>30000	40	метки на уровне изображения
LMD [23]	2019	5000	16	локальные патчи с метками
DMS [37]	2022	44560	56	семантические маски
LAS&T [48]	2025	650000	-	-
MatPredict [47]	2025	252	14	карты нормалей, альбедо, глубины и BRDF

2.3 Методы распознавания материалов объектов на изображениях

Самые первые методы распознавания материалов опираются на такие инструменты, как фильтры Габора, локальные бинарные шаблоны и другие текстурные фильтры для извлечения статистических характеристик поверхности [4]. Позже появились гибридные решения: например, Deep Filter Banks [7] (FV-CNN) объединяет свёрточные признаки из предобученной VGG-16, что позволяет достичь 75% точности на DTD.

Более современные решения часто используют известные архитектуры компьютерного зрения. Например, FCN + CRF [25] достигает 44% mIoU на MINC. DeepLabV3+ показывает до 42% mIoU на датасете Apple DMS. MateViT [44] объеди-

няет Vision Transformer с локальными декодерами для сегментации, достигая 51.1% mIoU на Apple DMS. В этой работе особенное внимание уделено быстродействию архитектуры. В работе [46] представляют ConCLIP – модифицированный CLIP с модулем объединения текстовых и картиночных признаков для задачи классификации строительных отходов (всего 5 классов).

Отдельный интерес представляет работа Multimodal material segmentation [35]. Авторы предлагают дополнить RGB-изображение поляризацией и инфракрасными картами, что помогает точнее определять характеристики поверхности объекта. В данной работе решается задача семантической сегментации, и авторы предлагают мультимодальную архитектуру на основе DeepLabV3+.

2.4 Формальная постановка задачи

В данной работе решается задача распознавания материалов, из которых сделаны объекты, по изображениям объектов. В общем случае эту задачу можно сформулировать как задачу классификации с несколькими метками.

Пусть существует датасет

$$\mathcal{D} = \left\{(x_i,\, \mathbf{m}_i)\right\}_{i=1}^{N}, \quad \mathbf{m}_i = (m_{i,1}, \ldots, m_{i,M})^\top \in \{0,1\}^M,$$

где $x_i \in \mathcal{X}$ — входное изображение i–го объекта, N — число объектов в датасете, M — число возможных материалов, а $\mathbf{m}_i$ — бинарный вектор длины M, в котором

$$m_{i,j} = \begin{cases} 1, & \text{если материал } j \text{ присутствует в объекте } i, \\ 0, & \text{иначе.} \end{cases}$$

Необходимо предоставить параметрическую модель

$$f(x;\theta)\colon \mathcal{X} \;\to\; [0,1]^M,$$

которая по входному изображению x возвращает вектор предсказаний присутствия каждого из M материалов. В данном случае θ – параметры модели.

Обычно выходной вектор такой модели рассматривают как набор вероятностей присутствия материалов и для получения финального бинарного вектора предсказаний применяют пороговую функцию. В данной работе будут также рассмотрены алгоритмы распознавания, в которых предсказания моделей будут переводиться в бинарный вид без применения пороговой функции (см. раздел 3.3).

Для обучения модели обычно используется усреднённая по всем материалам бинарная кросс-энтропия:

$$\mathcal{L}_{BCE}(\theta) = -\frac{1}{N\,M}\sum_{i=1}^{N}\sum_{j=1}^{M}\big[m_{i,j}\,\log\hat{y}_{i,j}\;+\;(1-m_{i,j})\,\log(1-\hat{y}_{i,j})\big]. \tag{2.1}$$

Эта функция потерь будет применяться в разделах 5.4, 5.5.

В разделе 5.1 рассмотрено обучение случайного леса на основе статистических признаков изображения, и в этом случае нет глобальной оптимизации на основе функции потерь: существует только локальная оптимизация разбиения в узлах деревьев, для которой применяется критерий Джини.

В разделе 3.2 задача определения материала сводится к задаче семантической сегментации. В таком случае происходит оптимизация кросс-энтропии:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{|\Omega|}\sum_{u\in\Omega}\sum_{c=1}^{C} y_{u,c}\,\log p_{u,c},$$

где Ω – множество всех пикселей в изображении, C – число классов, $y_{u,c} \in \{0,1\}$ – истинная метка класса c в пикселе u, $p_{u,c} \in [0,1]$ – предсказанная моделью вероятность принадлежность пикселя u к классу c.

В разделе 3.3 будет рассмотрен алгоритм инференса визуальных языковых моделей без дообучения.

Оценка качества классификации подробно рассматривается в разделе 3.1.3.

3 Метод

В данном разделе предложены следующие подходы для распознавания материалов объектов по их изображениям: классические методы на основе статистических признаков, семантическая сегментация, алгоритм на основе инференса визуальных языковых моделей, контрастивное обучение.

3.1 Определение материалов на основе статистических признаков признаков

Рассмотрим решение задачи классификации с несколькими метками материалов по изображениям объектов на основе статистических признаков. Алгоритм классификации состоит из трёх основных этапов: извлечение признаков заданных типов из изображения объекта, построение вектора признаков и обучение ансамбля классификаторов. На изображении 1 приведена схема алгоритма.

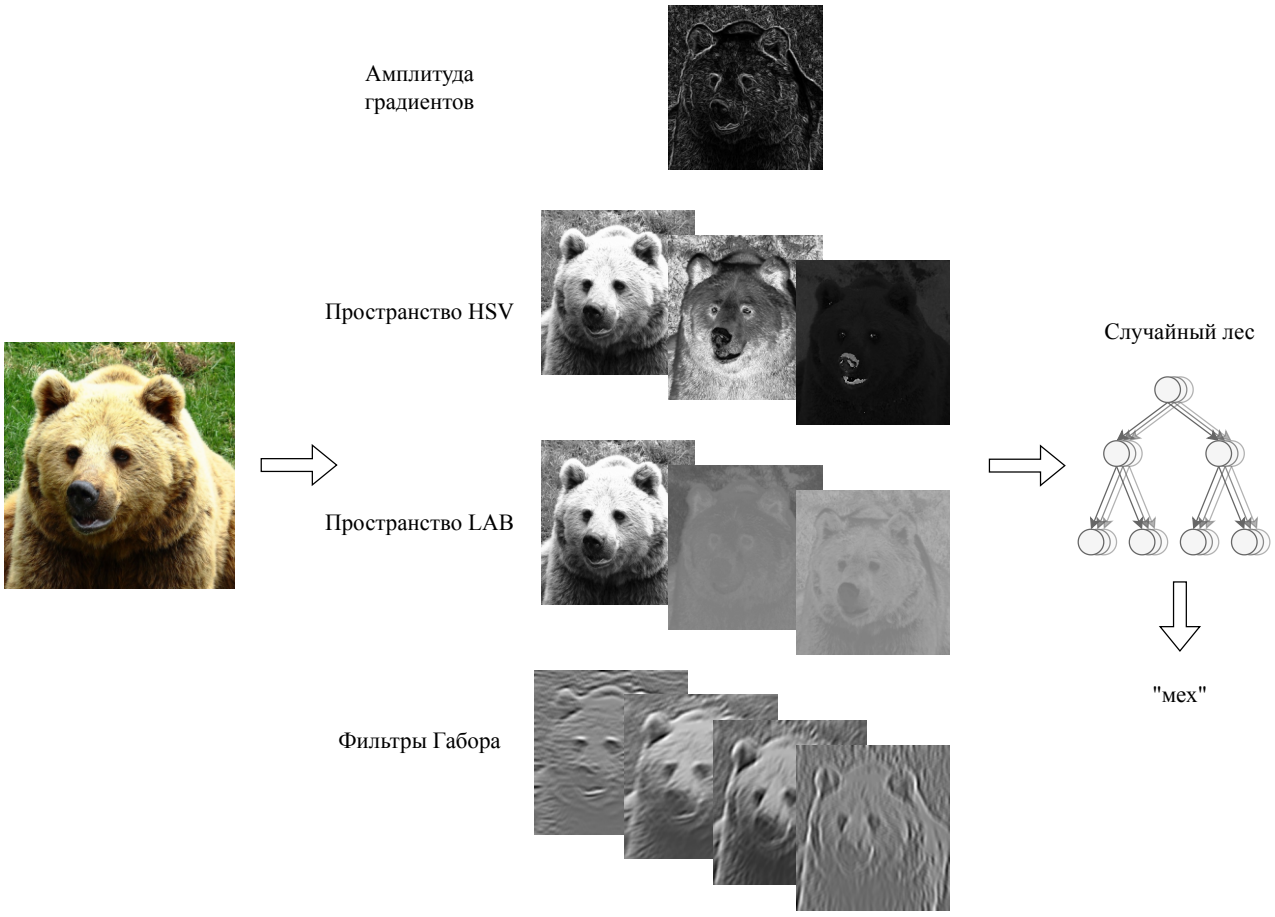


Рис. 1: Схема алгоритма классификации материала объекта на основе классических признаков.

3.1.1 Статистические признаки изображений

На этапе извлечения признаков из каждого входного изображения формируется совокупность описаний, включающая пять групп:

- **Признаки границ и краёв**, выявляющие локальные изменения яркости и структуры поверхностей;
- **Статистические признаки**, характеризующие распределение интенсивностей пикселей;
- **Признаки фильтров Габора**, отражающие текстурные свойства за счёт пространственно-частотных компонент;
- **Цветовые признаки**, описывающие пространственное распределение цветовых каналов;
- **Гистограмма LBP**, полезная для анализа структуры материала.

Рассмотрим подробнее каждую группу.

Признаки границ и краёв.

- **Резкость границ.** Этот параметр вычисляется как средняя величина градиента по всему избражению, которая в свою очередь вычисляется на основе фильтров Собеля [2]. Высокие значения резкости (>50) свойственны изображениям металлов с обработанными краями, стекла, пластика. Низкие значения (<20) свойственны изображениям тканей, меха и других мягких материалов.
- **Шероховатость текстуры.** Этот параметр вычисляется как обратная величина от дисперсии углов градиента, то есть на основе предыдущей величины. Высокие значения (близко к 1) свойственны изображениям гладких, глянцевых материалов: металл, пластик, стекло. Низкие значения (близко к 0) свойственны изображениям неоднородных текстур: бетон, грубая ткань, кора дерева.
- **Плотность границ.** Этот параметр отражает долю пикселей с сильными градиентами (больше среднего). Высокие значения (>0.5) свойственны изображениям материалов с явной текстурой, а низкие (<0.2) – изображениям однородных поверхностей.

Статистические признаки.

- **Эксцесс (Kurtosis)** рассчитывается по следующей формуле:

$$\text{kurtosis} = \frac{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^4}{\sigma^4 + \varepsilon} - 3,$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} I_i, \quad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^2},$$

  Где I_i – интенсивности пикселей, N – общее число пикселей, μ – среднее значение интенсивности, σ – стандартное отклонение, ε – малая константа для предотвращения деления на ноль. Эксцесс показывает, есть ли на изображении доминирующая яркость. Высокий эксцесс (>3): свойственен для однородных материалов одного цвета. Низкий эксцесс (<0) свойственен для неоднородных материалов, например, пестрой ткани.

- **Стандартное отклонение** показывает разброс значений яркости. Высокое (>50) свойственно для контрастных узоров, грубых текстур; низкое (<20) – для однородных поверхностей.

- **Энтропия.** Высокая энтропия (>6) также свойственна для сложных структур (органические структуры, ткани, природный камень), а низкая (<4) – для однотонных поверхностей.

- **Асимметрия (Skewness)**, или смещение распределения яркости. Вычисляется по формуле

$$\text{skewness} = \begin{cases} \dfrac{\frac{1}{N}\sum_{i=1}^{N}(I_i - \mu)^3}{\sigma^3}, & \sigma > \varepsilon, \\ 0, & \sigma \le \varepsilon, \end{cases}$$

  с теми же параметрами, что и эксцесс. Положительная асимметрия означает, что на изображении больше тёмных пикселей (тени, тёмные материалы), а отрицательная – что больше светлых (блестящие материалы).

- **Средняя интенсивность.** Помогает оценить материал, например белый пластик будет иметь значение около 200-250, а черная резина – около 20-50.

Фильтры Габора [1]. Набор фильтров Габора с ориентациями от 0° до 90° с шагом 15° применяется к изображению объекта. Каждая из приведенных ориентаций связана с извлечением текстурных элементов, вытянутых вдоль соответствующего направления:

- 0°: горизонтальные структуры (слоистая древесина, текстиль с продольным плетением);
- 15°, 30°: слегка наклоненные узоры, характерные для древесных волокон или вытянутых трещин на камне;
- 45°: диагональные штрихи – могут встречаться в ткани с «ёлочкой», на поверхности металла после шлифовки.
- 60°, 75°: обратный наклон: например, текстуры камня или текстиля с косой стежкой;
- 90°: вертикальные структуры (картина годичных колец дерева, трубчатая структура некоторых волокон).

После применения фильтров Габора для каждого изображения вычисляются следующие статистические показатели: среднее значение, стандартное отклонение и диапазон значений (разность между максимальным и минимальным значениями).

Цветовые признаки. Для получения цветовых признаков изображение конвертируется в пространства HSV и LAB (первая компонента отвечает за перцептивную яркость, вторая – за баланс красного и зеленого, третья – за баланс желтого и синего). Для каждого из каналов вычисляются среднее и стандартное отклонение.

Второй набор цветовых признаков – гистограммы для каждого из каналов исходного изображения (RGB). Строится 3 гистограммы по 16 бинов в каждой. Как правило, более узкое распределение свойственно однородным материалам, широкое – природным материалам и текстурированным поверхностям. Также возможны характерные для определенных материалов цветовые пики (например, изображению дерева соответствует распределение в коричневых тонах).

Гистограмма LBP (Local boundary pattern). LBP показывает локальные особенности текстуры изображения, фиксируя шаблоны градиентов яркости пикселей в его окрестности.

Алгоритм построения LBP следующий:

1. Каждый пиксель сранивается с 8 соседями;
2. Если сосед ярче, записываем на соответствующую позицию (0-7) единицу, иначе 0;
3. Таким образом получается 8-битный код, который переводится в диапазон значений 0-255;
4. Для полученного изображения строится гистограмма.

Эта гистограмма тоже помогает в понимании структуры материала: для гладких поверхностей будут доминировать коды 0 и 255, для текстур с воспроизводящимя маттерном будут присутствовать пики в определенных значениях гистограммы; для случайных текстур (например, бетон) распределение кодов будет равномерным.

Объединение признаков. Выделенные признаки объединяются в единый вектор признаков фиксированной размерности. В качестве классификатора используется алгоритм случайного леса [3], обладающий высокой устойчивостью к переобучению и способностью обрабатывать признаки различной природы. Обучение модели производится на датасете, рассмотренном в разделе 4.

3.1.2 Оценка важности признаков

В данном разделе на классических признаках изображения строится случайный лес, то есть ансамбль деревьев. Разделение в узлах деревьев происходит на основе критерия Джини:

$$G(v) = 1 - \sum_{k=1}^{K} p_k^2,$$

где p_k – доля класса k в узле. На основе изменения этой величины при переходе из узла-родителя к узлу-потомку можно определить важность того или иного класса в дереве:

$$VI_i^{(t)} = \sum_{v \in \mathcal{N}} \mathbb{1}(f(v) = i) \cdot p(v) \cdot \Delta_v,$$

где $\mathcal{N}$ – множество узлов дерева t, $f(v)$ – признак, по которому происходит разделение в узле v; $\Delta_v = p(v) \cdot G(v) - p(v_L) \cdot G(v_L) - p(v_R) \cdot G(v_R)$, v_L и v_R – левый и правый дочерние узлы.

Тогда для целого леса важность признака можно определить как среднюю важность признака по деревьям:

$$VI_i = \frac{1}{T} \sum_{t=1}^{T} VI_i^{(t)} \tag{3.1}$$

Для группы признаков G важность определяется как:

$$VI_G = \sum_{i \in G} VI_i \tag{3.2}$$

3.1.3 Метрики качества классификации с несколькими метками

Рассмотрим адаптацию стандартных метрик классификации для задачи классификации с несколькими метками.

Пусть есть N объектов и K возможных меток материалов. Обозначим для каждой метки $k \in \{1, \ldots, K\}$:

- TP_k – число верно предсказанных положительных примеров для метки k;
- FP_k – число ложноположительных предсказаний для метки k;
- FN_k – ложноотрицательных предсказаний для метки k.

Найдем точность (precision), полноту (recall) и F_1-меру (F_1-score) с микро- и макро-усреднением.

Точность модели отражает, какая доля предсказанных моделью меток является верной. Высокая точность означает, что модель редко ошибается (не добавляет лишних, неверных меток). Низкая точность говорит о большом числе ложноположительных срабатываний.

Полнота модели показывает, какую долю истинных меток предсказывает модель. Высокая полнота означает, что модель практически не пропускает нужные метки. Низкая полнота указывает на то, что модель консервативна и многие истинные метки остаются невыявленными.

F_1-мера – это гармоническое среднее точности и полноты. Она достигает высоких значений только если одновременно и точность, и полнота велики; эта метрика устойчива к дисбалансу классов.

Каждая метрика принимает значения из диапазона (0, 1).

Micro-precision, micro-recall и micro-F1. Просуммируем TP, FP, FN по всем меткам:

$$TP = \sum_{k=1}^{K} TP_k, \quad FP = \sum_{k=1}^{K} FP_k, \quad FN = \sum_{k=1}^{K} FN_k.$$

Найдем точность, полноту и F_1-меру:

$$\text{Precision}^{\text{micro}} = \frac{TP}{TP + FP}, \quad \text{Recall}^{\text{micro}} = \frac{TP}{TP + FN}, \tag{3.3}$$

$$\text{F}_1^{\text{micro}} = 2 \cdot \frac{\text{Precision}^{\text{micro}} \cdot \text{Recall}^{\text{micro}}}{\text{Precision}^{\text{micro}} + \text{Recall}^{\text{micro}}}. \tag{3.4}$$

Эти метрики "взвешивают" все метки одинаково, независимо от их частоты.

Macro-precision, macro-recall и macro-F_1. Найдем precision и recall для каждой метки:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}.$$

Теперь усредним по всем классам:

$$\text{Precision}^{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \text{Precision}_k, \quad \text{Recall}^{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} \text{Recall}_k, \tag{3.5}$$

$$\text{F}_1^{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{3.6}$$

Каждая метка имеет одинаковый вес, даже если она встречается редко. Низкое значение macro-F_1 может указывать на то, что модель плохо работает на менее частых классах.

Subset Accuracy. Эта метрика также принимает значения от 0 до 1. Она показывает долю объектов, для которых модель верно предсказала все метки.

$$\text{Subset Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\hat{Y}_i = Y_i\}, \tag{3.7}$$

где $Y_i \subseteq \{1, \ldots, K\}$ — истинный набор меток для объекта i, а $\hat{Y}_i$ — предсказанный моделью набор.

3.2 Определение материалов с применением семантической сегментации

Рассмотрим задачу определения материала как классическую задачу семантической сегментации. Её частный случай, задача активного восприятия сцены, встречается в робототехнике [52]. В разделе 5.5 предлагается алгоритм обобщения этого подхода для поставленной изначально задачи многометочной классификации. Пусть есть $\mathcal{X}$ — множество входных изображений, и $\mathcal{Y} = \{0, 1\}^{H \times W \times K}$, где H, W – размеры изображения, а K – число классов. $Y_{i,j,k} = 1$ означает, что пиксель (i, j) принадлежит классу k. Пусть дан обучающий набор данных $\mathcal{D} = \{(x_n, y_n)\}_{n=1}^{N}, \quad x_n \in \mathcal{X},\ y_n \in \mathcal{Y}$. Тогда необходимо найти отображение f^*:

$$f^* : \mathcal{X} \to \mathcal{Y}, \quad \forall (x_n, y_n) \in \mathcal{D} : \quad f^*(x_n) = y_n.$$

3.2.1 Модели семантической сегментации

В качестве $\mathcal{D}$ рассмотрим датасет DMS. Это единственный датасет с полноценными семантическими масками, который размечен достаточно большим числом категорий. Предлагается обучить сегментаторы HRNet, DDRNet и Mask2Former.

High-Resolution Network (HRNet) [20] (2019) – это сверточная архитекура, которая сохраняет высокое пространственное разрешение на всём протяжении обработки. Архитектура состоит из параллельно работающих "ветвей" разного разрешения, между которыми после каждого блока происходит двусторонний обмен информацией. Такой дизайн позволяет одновременно извлекать детализированные локальные признаки и контекстные представления, что повышает точность выделения мелких объектов и тонких границ.

DDRNet [29] (2022) развивает идею, предложенную в HRNet, оптимизируя её для реального времени за счёт сокращения числа ветвей до двух: одной с высоким и одной с низким разрешением. Ключевым элементом выступают "диагональные" блоки обмена, которые обеспечивают более эффективную передачу контекста между картами признаков с низким разрешением в карты с высоким разрешением.

Mask2Former [28] (2021) – это универсальный подход, объединяющий инстанс-сегментацию (то есть сегментацию отдельных объектов), семантическую и паноптическую сегментацию в единую архитектуру на основе трансформеров.

3.2.2 Оценка качества моделей

Для оценки качества работы моделей воспользуемся классическими метриками семантической сегментации: Mean Intersection over Union (mIoU) и Mean Average Precision (mAP).

Пусть K – количество классов в задаче семантической сегментации. Обозначим для каждого класса k множество пикселей, предсказанных как k, через P_k, а истинных пикселей класса k – через G_k.

Mean Intersection over Union (mIoU). Эта метрика принимает значения из диапазона [0, 1] и показывает долю перекрытия предсказанных и истинных областей по всем классам $k \in K$:

$$\mathrm{mIoU} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{IoU}_k. \tag{3.8}$$

Здесь

$$\mathrm{IoU}_k = \frac{|P_k \cap G_k|}{|P_k \cup G_k|},$$

где $|\cdot|$ означает число пикселей в множестве.

Mean Average Precision (mAP). Для каждого класса k строится кривая Precision–Recall путём изменения порога уверенности модели. Обозначим функцию точности как

$$P_k(r), \quad r \in [0, 1]$$

— точность при полноте r. Тогда средняя точность (average precision) для класса k определяется как площадь под этой кривой:

$$\mathrm{AP}_k = \int_0^1 P_k(r)\, dr.$$

И, наконец,

$$\mathrm{mAP} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{AP}_k. \tag{3.9}$$

Эта метрика тоже принимает значения в диапазоне от 0 до 1. Она показывает, насколько хорошо модель предсказывает класс пикселя, соблюдая баланс между точностью и полнотой.

3.3 Определение материалов с помощью визуально-языковых моделей

Визуально-языковые модели на основе больших языковых моделей сочетают фундаментальные знания о мире и возможность обрабатывать изображения. Как правило, такие модели состоят из большой языковой модели и энкодера изображения, который строит представление изображения в виде токенов. Традиционные задачи при обучении VLM – это ответы на вопросы по изображению (VQA) и генерация описаний. Первая задача имеет место в данном разделе: предлагается построить классификацию материалов объектов на изображении на основе инференса визуально-языковых моделей. Вторая задача обсуждается в разделе 4.

Итак, необходимо решить задачу классификации с несколькими метками. Алгоритм определения материала состоит из следующих этапов: обрезка изображения с небольшим отступом по границам маски объекта; инференс VLM; постобработка ответа модели.

Предлагается сравнить 2 модели: QWEN2.5-VL [45] и LLaVa1.5 [42], обе в размерах 7B.

QWEN2.5-VL и LLaVA-1.5 отличаются прежде всего своими визуальными энкодерами и масштабом данных: QWEN2.5-VL использует собственный ViT [19] с динамическим разрешением, что позволяет обрабатывать мелкие детали и работать с изображениями разных размеров. Модель обучалась примерно на 1.4 млрд пар "изображение-текст" и затем дообучалась на мультизадачных высококачественных аннотациях: VQA, OCR (распознавание символов на изображениях). Финальный этап обучения – тонкая настройка для режима диалога.

В LLaVA-1.5 использован обычный энкодер ViT-L/14 с фиксированным входным разрешением 336×336 и MLP-проекцией визуальных признаков в пространство эмбеддингов языковой модели Vicuna [41]. LLaVa-1.5 обучалась примерно на 595 тыс. пар "изображение-текст" из датасета CC3M (фильтрованная выборка).

Для оценки качества работы моделей будем использовать набор метрик классификации из раздела 3.1.3. Оценку будем производить на основе датасета, который

обсуждается в разделе 4.

3.4 Контрастивное обучение мультимодальной архитектуры

Этот раздел работы, наравне с разделом 4, является центральным. Ранее было показано, что существующие датасеты и подходы достаточно сильно ограничены в разнообразии категорий, причем этот недостаток свойственен для любых архитектур с закрытым словарем. В разделе 3.3 рассматривается решение на основе VLM, в котором словарь не является закрытым. Это хорошее решение, однако оно плохо подходит для задач в реальном времени: время работы даже небольшой VLM обычно исчисляется сотняти микросекунд. В данном разделе предлагается рассмотреть решение с открытым словарем на основе контрастивного обучения "изображение-текст": обычно в таких архитектурах применяются легковесные и быстрые модели.

Рассмотрим датасет

$$\mathcal{D} = \{(x_i,\, c_i,\, m_{i,j},\, T_i)\}_{i=1}^{N}{}_{j=1}^{M},$$

где x_i — изображение объекта, c_i — его семантический класс, $m_{i,j}$ — материал или набор материалов, из которых сделан объект, а t_i — текстовое описание для этого объекта.

Введем два отображения:

$$f : \mathcal{X} \to \mathbb{R}^d, \quad g : \mathcal{L} \to \mathbb{R}^d,$$

где $f(x)$ — векторное представление изображения объекта, а $g(t)$ — векторное представление соответствующего описания, существующие в общем d-мерном пространстве.

Обучение моделей сводится к решению оптимизационной задачи

$$\min_{f,g} \mathcal{L}(f, g;\, \mathcal{D}),$$

где $\mathcal{L}$ — общая контрастивная функция потерь, которая должна обеспечить:

- согласование эмбеддингов $f(x_i)$ и $g(t_i)$ для одного объекта;
- разделение эмбеддингов представителей разных категорий объектов;
- разделение эмбеддингов по материалам;

- сохранение семантической структуры данных.

Построение эмбеддингов объектов позволяет говорить о непрерывном распределении признаков, содержащихся в них. Фокус на сближении эмбеддингов описаний и эмбеддингов изображений объектов позволит более эффективно выделять визуальные признаки объектов. Полученные эмбеддинги можно использовать для решения downstream задач: например, для построения классификатора поверх таких эмбеддингов. О функции потерь и оценке качества модели подробно рассказано далее.

Цель контрастивного обучения заключается в построении латентных представлений исходных данных, которые были бы близки для схожих объектов и отличались для остальных. Обычно под сходством понимают близость эмбеддингов, и чаще всего это косинусная близость:

$$\text{sim}(z_i, z_k) = \frac{z_i^\top z_k}{\|z_i\| \, \|z_k\|}, \tag{3.10}$$

где $||\cdot||$ обозначает евклидову норму вектора.

В контрастивном обучении применяют как обучение без учителя, так и обучение с учителем.

Контрастивное обучение без учителя. Как правило, этот алгоритм строится на аугментации исходных данных: например, в случае изображений сближают эмбеддинги для фрагментов одной и той же картинки. Примеры таких работ – CPC [15], в котором обучают автоэнкодер на различных данных (текст, аудио, изображения) и SimCLR [18], в котором упор сделан на аугментации и большой размер батча и рассматриваются только изображения.

Контрастивное обучение с учителем. В обучении с учителем ключевую роль играет разметка: эмбеддинги элементов одного класса стараются сблизить, а разных классов – отдалить. Этот подход полезен в данной работе, поскольку одна из поставленных целей – построение эмбеддингов, содержащих информацию о материале. Один из наиболее известных примеров контрастивного обучения – SupCon [24]. В этой работе обучают стандартную в компьютерном зрении архитектуру: сверточный энкодер и проекция на основе MLP. Функция потерь задается следующим образом: пусть есть батч из N изображений $\{x_i\}$ и меток $\{y_i\}$, а $z_i = f(x_i) \in \mathbb{R}^d$ – проекции изображений. Пусть множество положительных примеров для i-го эле-

мента $\mathcal{P}(i) = \{p \neq i : y_p = y_i\}$, а τ – температура, которая "контролирует" резкость распределения. Тогда функция потерь

$$\mathcal{L}_{\text{SupCon}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp\big(\text{sim}(z_i,z_p)/\tau\big)}{\sum\limits_{k=1}^{N}\mathbb{1}_{[k\neq i]}\,\exp\big(\text{sim}(z_i,z_k)/\tau\big)}\right] \tag{3.11}$$

Эта функция потерь будет далее модифицирована: как будет показано, датасет, на котором обучаются модели, содержит большое число уникальных материалов (около 1600), часть из которых синонимична друг другу, например металл и сталь. Поскольку один объект может состоять из нескольких материалов, при построении эмбеддингов предпочтительно учитывать каждый из них. Посчитать сходство материалов можно по косинусной близости их текстовых эмбеддингов, построенных с помощью текстовой модели. Желательно не переобучить модель под категории материалов, то есть свести задачу к обычной классификации, поэтому будем учитывать также категорию объекта.

Пусть снова есть батч из N изображений $\{x_i\}$ и меток $\{y_i\}$, а $z_i = f(x_i) \in \mathbb{R}^d$ – проекции изображений. Пусть также имеется $S \in \mathbb{R}^{C\times C}$ – матрица схожести между метками классов. Пусть $M \in \{0,1\}^{N\times C}$ – матрица, состоящая из multi-hot векторов, которые кодируют эталонные классы для объектов.

Определим матрицу сходства $T \in \mathbb{R}^{N\times N}$ на основе меток следующим образом:

$$T_{ij} = \frac{(MSM^\top)_{ij}}{\max\Big((\sum_{c=1}^{C} M_{ic})(\sum_{c=1}^{C} M_{jc}),\ 1\Big)} \tag{3.12}$$

Пусть τ – некоторое пороговое значение, определяющее сходство материалов. Определим множество положительных примеров для i-го элемента следующим образом:

$$\mathcal{P}'(i) = \{j \in \{1,\dots,N\} \setminus \{i\} \mid T_{ij} \geqslant \tau_{\text{mat}}\}$$

Тогда новая функция потерь будет иметь вид:

$$\mathcal{L}_{\text{MultiSupCon}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{1}{|\mathcal{P}'(i)|}\sum_{p\in\mathcal{P}'(i)}\log\frac{\exp\big(\text{sim}(z_i,z_p)/\tau\big)}{\sum\limits_{k=1}^{N}\mathbb{1}_{[k\neq i]}\exp\big(\text{sim}(z_i,z_p)/\tau\big)}\right] \tag{3.13}$$

Функция потерь MultiSupCon далее будет использована как компонента общей функции потерь.

Контрастивное обучение "текст-изображение". Цель данного метода состоит в сбижении эмбеддингов для изображений и соответствующих им описаний. Архитектура модели представляет параллельное применение энкодера изображения и текстового энкодера. Наиболее известный представитель такого класса моделей – CLIP [30].

Для обучения такой модели в рамках поставленной задачи необходим достаточно большой датасет с изображениями объектов и описаниями на естественном языке, содержащими материалы объектов. В разделе 2.2 показано, что на данный момент не существует датасетов, специализированных на описании материалов объектов. В данной работе представлен подходящий датасет и его создание подробно рассмотрено в разделе 4.

Функция потерь, которая используется при обучении CLIP, также известна как InfoNCE (Information Noise-Contrastive Estimation). Она вводится следующим образом: пусть $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ – батч L_2-нормализованных эмбеддингов изображений, $\{y_j\}_{j=1}^N \subset \mathbb{R}^d$ – батч L_2-нормализованных текстовых эмбеддингов, а их матрица сходства (similarity) получена перемножением:

$$S_{ij} \;=\; \frac{1}{\tau}\, x_i^\top y_j, \tag{3.14}$$

где $\tau = 1/\exp(\text{logit scale})$ – температура, которая задана через обучаемый параметр logit scale. Тогда кросс-энтропия "изображения $\to$ "текст" имеет вид:

$$\mathcal{L}_{\text{i}\to\text{t}} = -\frac{1}{N}\sum_{i=1}^{N} \log\!\left(\frac{\exp\big(S_{ii}\big)}{\sum_{j=1}^{N}\exp\big(S_{ij}\big)}\right),$$

а кросс-энтропия "текст $\to$ "изображения" дается выражением:

$$\mathcal{L}_{\text{t}\to\text{i}} = -\frac{1}{N}\sum_{i=1}^{N} \log\!\left(\frac{\exp\big(S_{ii}\big)}{\sum_{j=1}^{N}\exp\big(S_{ji}\big)}\right).$$

Итоговая функция потерь является усреднением двух предыдущих функций:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{2}\left(\mathcal{L}_{\text{i}\to\text{t}} + \mathcal{L}_{\text{t}\to\text{i}}\right) \tag{3.15}$$

3.4.1 Общая функция потерь

Согласно постановке задачи кажется уместной комбинация функций потерь (3.13) и (3.15): первая компонента обеспечит сближение для эмбеддингов синонимичных материалов и отдалит остальные эмбеддинги, а вторая компонента сблизит текстовые и эмбеддинги изображений, то есть поможет выделить больше информации о контексте.

3.4.2 Оценка качества работы модели

Основная цель данной части работы состоит в обучении модели компьютерного зрения, которая будет строить информативные эмбеддинги для изображений объектов. Такие эмбеддинги должны содержать информацию как о категории, так и о материале объекта. При этом можно оценивать как непосредственно сами эмбеддинги, так и способность модели решать zero-shot задачи. Оценка модели складывается из следующих этапов:

- Оценка согласованности между эмбеддингами изображений объектов и эмбеддингами их описаний: для этого применяют такие метрики, как MCS (Mean Cosine Similarity), MRR (Mean Reciprocal Rank) и Recall@K. Эти метрики также применяют во время обучения модели для отбора лучших весов.

- Оценка согласованности схожести эмбеддингов изображений объектов и меток для категорий материалов: корреляция Пирсона, коэффициент силуэта кластера, скорректированный индекс Рэнда, классификация на основе k-means.

- Оценка применимости эмбеддингов для решения более общих задач: например, обучение небольшого классификатора для предсказания меток классов, в нашем случае материалов.

Входные параметры следующие:

- $E \in \mathbb{R}^{N \times D}$ – предсказанные эмбеддинги N изображений или описаний;

- $C \in \{0,1\}^{N \times N_{\text{cat}}}$ – матрица категорий материалов (N one-hot векторов);

- $M \in \{0,1\}^{M \times M_{\text{mat}}}$ – матрица категорий объектов (N multi-hot векторов: один объект может состоять из нескольких материалов);

- $S_{\text{cat}} \in \mathbb{R}^{N_{\text{cat}} \times N_{\text{cat}}}$ – матрица схожести между категориями объектов;
- $S_{\text{mat}} \in \mathbb{R}^{N_{\text{mat}} \times N_{\text{mat}}}$ – матрица схожести между материалами.

Рассмотрим подробнее перечисленные метрики и начнем с оценки согласованности между эмбеддингами изображений и соответствующих описаний.

Пусть имеется $\{x_i\}_{i=1}^N \subset \mathbb{R}^d$ – батч L_2 нормализованных эмбеддингов изображений и $\{y_j\}_{j=1}^N \subset \mathbb{R}^d$ – батч L_2 нормализованных текстовых эмбеддингов, а также их матрица схожести S (3.14).

Пусть для каждого изображения i, существует r_i – позиция соответствующего ему описания среди всех N описаний:

$$r_i \;=\; \big|\{\, j : S_{ij} \geq S_{i,i}\}\big|. \tag{3.16}$$

Средняя косинусная близость (MCS) показывает среднюю меру сходства между эмбеддингами изображений и их корректными текстовыми описаниями: чем выше MCS, тем более согласованы пространства изображений и текстов.

$$\text{MCS} \;=\; \frac{1}{N}\sum_{i=1}^{N} S_{ii}. \tag{3.17}$$

Средний обратный ранг (MRR) показывает, насколько высоко в среднем оказывается корректное текстовое описание объекта при ранжировании всех описаний по убыванию косинусной близости к эмбеддингу изображения. Чем выше значение MRR, тем ближе к первой позиции находится правильное описание, то есть тем лучше модель сопоставляет изображения и тексты.

$$\text{MRR} \;=\; \frac{1}{N}\sum_{i=1}^{N} \frac{1}{r_i}. \tag{3.18}$$

Recall@K показывает долю случаев, в которых корректное текстовое описание объекта оказалось среди первых K по косинусной близости к эмбеддингу изображения. Высокое значение Recall@K означает, что правильные описания попадают в топ-K рекомендаций.

$$\text{Recall@}K \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[r_i \leq K\big], \tag{3.19}$$

где $\mathbf{1}[\cdot]$ – индикаторная функция.

Как правило, MRR и Recall@K достаточно сильно скоррелированы. Заметим, что InfoNCE максимизирует именно эти метрики, в то время как значение MCS может падать одновременно с ростом MRR и Recall@K. Ещё одна особенность приведенных метрик – их зависимость от размера батча, поэтому при сравнении моделей важно сохранять неизменным этот параметр.

Теперь рассмотрим оценку самих эмбеддингов. Согласно постановке задачи, в первую очередь нас интересуют эмбеддинги изображений объектов, хотя аналогичный подход можно распространить и на текстовые эмбеддинги.

Начнем с "внутренних" метрик для эмбеддингов: корреляция, метрики кластеризации. Целевые матрицы сходства между эмбеддингами строятся по категориям объектов и материалам так же, как и в случае функции SupCon, согласно формуле 3.12. Воспользуемся C и S_{cat} для получения T_{cat}, а M и S_{mat} для получения T_{mat}. Построим матрицу сходства эмбеддингов S, воспользовавшись формулой 3.14.

Корреляция Пирсона характеризует линейную связь между вектором предсказанных сходств эмбеддингов $s = (s_i)$ и вектором истинных сходств $t = (t_i)$:

$$\text{Pearson}(s,t) = \frac{\text{Cov}(s,t)}{\sigma_s \cdot \sigma_t} = \frac{\sum_{i=1}(s_i - \overline{s})(t_i - \overline{t})}{\sqrt{\sum_{i=1}(s_i - \overline{s})^2}\sqrt{\sum_{i=1}(t_i - \overline{t})^2}}$$

Значение метрики лежит в диапазоне $[-1, 1]$:

- $+1$ — полная прямая линейная зависимость;
- -1 — полная обратная линейная зависимость;
- 0 — отсутствие линейной связи.

Silhouette score измеряет качество разбиения данных на кластеры согласно меткам классов (в нашем случае материалов), отражая, насколько внутрикластерные расстояния малы относительно межкластерных. Значения метрики лежат в $[-1, 1]$:

- $+1$ — объекты хорошо отделены в пределах своих кластеров;
- 0 — кластеры пересекаются или нечетко выражены;
- -1 — объекты могут быть ошибочно отнесены к соседним кластерам.

Найдем $a(e_i)$, средним расстояние d от i-го эмбеддинга до других эмбеддингов того же класса (то есть эмбеддингов из кластера C_i) и $b(e_i)$ – минимум по всем остальным кластерам среднего расстояния от i-го эмбеддинга до других эмбеддингов в этом кластере:

$$a(e_i) \;=\; \frac{1}{|C_i|-1}\sum_{\substack{j\in C_i\\ j\neq i}} d(e_i,e_j), \qquad b(e_i) \;=\; \min_{C\neq C_i}\frac{1}{|C|}\sum_{j\in C} d(e_i,e_j),$$

Значение silhouette score задается следующей формулой:

$$\text{Silhouette} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{b(i)-a(i)}{\max\{a(i),\,b(i)\}}. \tag{3.20}$$

Классификация на основе k-средних проверяет, насколько хорошо эмбеддинги группируются в соответствии с заданными метками (категориями объектов или материалами).

1. Разбиваем валидационную выборку эмбеддингов на тренировочную и тестовую части;

2. Обучаем алгоритм k-means на тренировочных эмбеддингах с k, равным числу реальных классов;

3. Применяем полученные центры кластеров для предсказания меток тестовых эмбеддингов;

4. Вычисляем стандартные классификационные метрики (accuracy, precision, recall, F_1) согласно разделу 3.1.3, чтобы оценить качество кластеризации как классификатора.

Mean Average Precision (mAP) оценивает качество ранжирования эмбеддингов изображений объектов по сходству с точки зрения того, насколько высоко в списке оказываются эмбеддинги одной категории. Для каждого тестового эмбеддинга $\mathbf{e}_i$ вводится бинарная релевантность

$$\mathrm{rel}_{ij} = \begin{cases} 1, & c_j = c_i, \\ 0, & \text{иначе}, \end{cases}$$

где c_i и c_j — истинные классы эмбеддингов $\mathbf{e}_i$ и $\mathbf{e}_j$. Отсортировав все тестовые эмбеддинги по убыванию сходства с $\mathbf{e}_i$, определяем

$$P_i(k) = \frac{1}{k} \sum_{j=1}^{k} \mathrm{rel}_{i,\pi_i(j)},$$

где $\pi_i(j)$ — индекс j-го по сходству эмбеддинга. Тогда Average Precision для $\mathbf{e}_i$ задаётся

$$\mathrm{AP}_i = \sum_{k=1}^{N_{\mathrm{tr}}} P_i(k)\, \Delta\mathrm{rel}_i(k), \quad \Delta\mathrm{rel}_i(k) = \mathrm{rel}_{i,\pi_i(k)} - \mathrm{rel}_{i,\pi_i(k-1)},$$

при $\mathrm{rel}_{i,\pi_i(0)} \equiv 0$. Наконец, mAP — это среднее AP по всем N_{te} тестовым эмбеддингам:

$$\mathrm{mAP} \;=\; \frac{1}{N_{\mathrm{te}}} \sum_{i=1}^{N_{\mathrm{te}}} \mathrm{AP}_i. \tag{3.21}$$

Валидация с MLP-классификатором материалов показывает, насколько информативны обученные эмбеддинги изображений объектов для задачи многометочной классификации материалов. В качестве классификатора используется небольшой MLP, обученный на валидационной части эмбеддингов, где каждому объекту соответствует от 1 до 5 материалов. Для оценки качества используются стандартные метрики многометочной классификации из раздела 3.1.3.

3.5 Архитектура модели

Поскольку цель данного раздела работы состоит в построении эмбеддингов для объектов на изображении, объекты необходимо локализовать. Интуитивно понятный способ сделать это состоит в обрезании изображения по границам маски или окружающего прямоугольника (bounding box), возможно с добавлением отступов. При этом часто встречаются объекты сложной формы, и, чтобы отделить их от фона, приходится прибегать к его размытию. Другие способы, такие как закраска фона, убирают контекст сцены и таким образом усложняют задачу. Отметим, что глобальный контекст часто может быть полезен для определения качеств объекта: например, стул, который стоит на улице, скорее будет пластиковым, а тот, что стоит в помещении – деревянным.

В данной работе предлагается сравнить обычный CLIP и модифицированный через добавление модуля, в котором происходит объединение эмбеддингов CLIP и эмбеддингов для всего изображения. На рисунке 2 показана схема модели: модуль

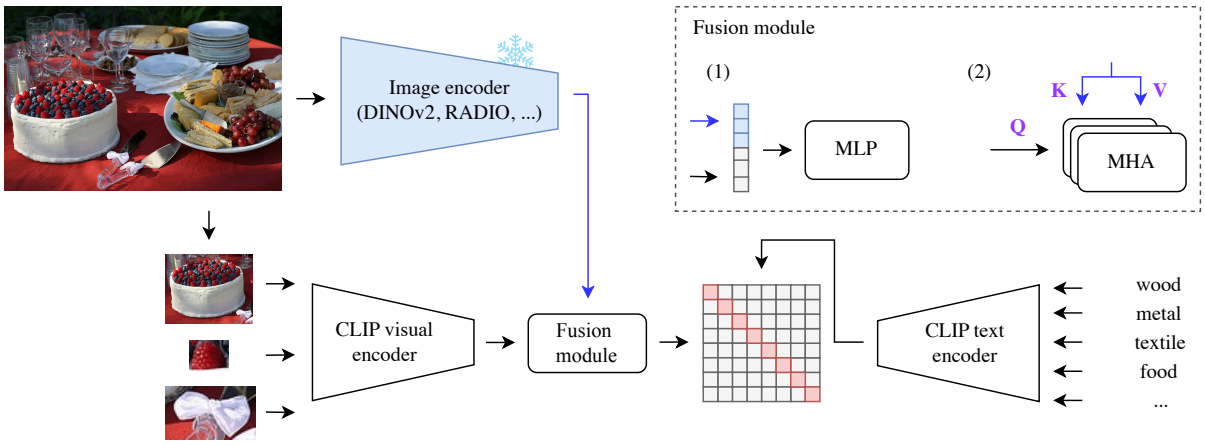


Рис. 2

для объединения эмбеддингов может быть обычным многослойным перцептроном (MLP) или более сложным блоком Multi-head attention (MHA) [13].

Объединение на основе MLP необходимо в качестве первоначального приближения: если этот эксперимент показывает положительные результаты (улучшение метрик по сравнению с обычным CLIP), то можно будет переходить к более сложному объединению признаков.

В качестве моделей для кодирования глобального контекста рассмотрены фундаментальные трансормерные модели DINOv2 [38] и RADIOv2.5 [51].

4 Разработка датасета MatLVIS

В данной работе реализована автоматическая доразметка датасета LVIS [17]. Методы автоматической разметки датасетов для задач компьютерного зрения также рассматриваются в работе [55]. LVIS содержит достаточно большое количество изображений, объекты на которых размечены бинарными масками; в датасете около 1200 уникальных категорий объектов. Всего в датасете более 1.2 млн объектов.

Сформулируем задачу следующим образом: пусть существует исходный датасет

$$\mathcal{D} = \left\{ \left(x_i,\ c_{i,j}\right)\right\}_{i=1\ j=1}^{N\ M}$$

где $\mathcal{X} = \{x_i\}_{i=1}^{N}$ – множество изображений с выделенными объектными (бинарными) масками, а $\{c_{i,j}\}_{i=1\ j=1}^{N\ M}$ – метки классов (N – число изображений, M – число объектов на каждом изображении). Необходимо дополнить этот датасет текстовым корпусом

$$\mathcal{T} = \{\, t_{i,j} \mid i = 1, \ldots, N,\ j = 1, \ldots, M\},$$

где для каждого объекта $x_{i,j}$ генерируется описание $t_{i,j}$ на естественном языке. Каждое описание должно удовлетворять следующим свойствам:

- описание должно включать класс объекта c_i и синонимы и материал m_i;
- текстовый корпус $\{t_{i,j}\}_{i=1\ j=1}^{N\ M}$ должен иметь достаточную лексическую и синтаксическую вариативность.

Помимо этого, требуется выделить $\{m_{i,j,k}\}_{i=1\ j=1\ k=1}^{N\ M\ K}$ – набор материалов для каждого элемента датасета (K – число материалов в описании).

В результате должен быть получен расширенный датасет

$$\widetilde{\mathcal{D}} = \left\{ \left(x_i,\ c_{i,j},\ t_{i,j},\ m_{i,j,k}\right)\right\}_{i=1\ j=1\ k=1}^{N\ M\ K},$$

в котором для каждого изображения с выделенным масками набором объектов также существует набор описаний и выделенных материалов.

Отметим, что некоторые категории в LVIS не могут быть ассоциированы с материалами: люди, животные, еда, жидкость и так далее. Эти категории имеют соответствующую метку и объединены таким образом в суперклассы.

4.1 Общее описание подхода к подготовке набора данных

На рисунке 3 приведена схема разметки датасета. Этот процесс разделен на два этапа: разметка существующих объектов описаниями и выделение упомянутых материалов. Рисунок 4 показывает первый этап разметки на примере двух изображений из датасета. Разметка происходит в автоматическом режиме с применением LLM и VLM. Первый этап разделяется на генерацию описания с помощью VLM и валидацию полученного описания с помощью LLM, а также доразметку объектов, которые получили некачественные описания. Второй этап реализован с применением LLM как для выделения материалов, так и для их валидации. В последующих разделах разметка датасета рассмотрена подробно.



Рис. 3: Схема разметки датасета: (1) Разметка изображений из датасета LVIS на основе существующих масок и категорий объектов; (2) Определение материалов, упомянутых в сгенерированных описаниях.

4.2 Разметка с применением VLM

Наиболее важная, но объемная по времени часть разметки датасета – генерация начальных описаний объектов. В данной работе особенно важно, чтобы описания содержали названия материалов, из которых сделан объект. При этом важно, чтобы словарь материалов был наиболее разнообразным и естественным.

Использование предобученных моделей с закрытым словарем плохо подходит для получения такой разметки. Во-первых, словарь ограничен небольшим числом материалов и включает не больше 60 категорий. Во-вторых, разметка сильно зависит от качества модели, которое, как будет показано в разделе 5, может быть недо-

статочным. Наконец, такие модели предсказывают только категорию материала, то есть полное описание необходимо генерировать на основе дополнительных данных, таких как категория объекта. Подобные описания не выглядят естественно и обычно не отличаются разнообразием, к тому же они не могут учитывать контекст всего изображения.

Для получения качественных описаний хорошо подходят VLM: они учитывают всё изображение и при этом используют фундаментальные знания о мире при генерации ответов. Выбор модели обусловлен ограничениями по вычислительным ресурсам и требованиями к качеству работы: в данном случае была выбрана модель QWEN2.5-VL [45] в размере 7B (8.29 миллиардов параметров). В разделе 3.3 подробнее рассказано об этой модели.

Процесс разметки состоит из подготовки изображения и генерации запроса для модели. Поскольку для каждого объекта в датасете требуется отдельное описание, необходимо локализовать объекты на изображении. В решении этой задачи отлично показал себя подход "Set of Marks" (набор марок) [40]. Основная идея подхода состоит в том, чтобы пронумеровать объекты и обвести контуры разными цветами. В таком случае модели проще понять, к какому объекту относится запрос; кроме того, сгенерированные ответы проще обработать. На изображении 4 (1) приведены примеры запросов, которые были использованы для разметки и ответы модели. Разметка одного изображения имеет форму диалога. Системная инструкция сформулирована следующим образом:

> I have labelled a bright numeric ID and a bright outline for some visual objects in the image. When I ask to provide a caption, pay **special attention to the material** of which the object is made. Do not tell anything about outline or its color. Fit each description in 10 words. Do you get the task? Answer only yes/no.

Первое предложение отвечает за атрибуты, на которые необходимо обратить внимание: эта формулировка также была использована авторами подхода SoM. Второе предложение отвечает за формулировку цели: описание объекта должно содержать материал. Более строгие формулировки приводят к ухудшению разнообразия описаний и одновременно увеличению галлюцинаций. Инструкция по исключению из описания обводки и номера объекта необходима, чтобы понизить число соответствующих случаев: без этой инструкции ответы часто содержат лишнюю информацию,

которая может ухудшить результаты будущего обучения. Инструкция, ограничивающая число слов, также уменьшает число галлюцинаций или просто описаний избыточной длины. Последняя инструкция нужна для предотвращения длинного ответа на приведенное сообщение, которое является началом диалога.

В таблицах 2 и 3 приведено количество качественных и некорректных сгенерированных описаний. Корректность описания определяется автоматически с применением LLM: подробнее об этом рассказано в следующем разделе. Из приведенных таблиц следует, что около половины всех размеченных объектов размечены качественно: остальные описания требуют доработки или переразметки.

Качественные описания содержат категорию объекта либо синоним и материал, они также часто включают цвет и другие качества объекта. При этом дополнительная информация иногда похожа на галлюцинации: например, кот на правой фотографии 4 имеет описание "Серо-белый кот с длинным хвостом". В данном описании нет материала и это корректно, однако исходя из изображения ничего нельзя сказать про хвост кота. Можно выделить несколько воспроизводящихся паттернов, которые имеют место в некорректных ответах:

- Упоминание цвета и размера объекта, но не материала;
- Упоминание в различных формулировках номера, присвоенного объекту при предобработке (пример можно увидеть в описании автобуса номер 7 на левом изображении 4);
- Описание самого номера на изображении: эта ошибка характерна для маленьких объектов, которые целиком закрыты номером.

Первая группа ошибок может быть исправлена автоматически на стадии валидации. Вторая группа ошибок частично исправляется посредством применения регулярных выражений и вырезания таких фрагментов из описания. Третья группа поддается исправлению только через повторную генерацию описания.

4.3 Валидация разметки с применением LLM

Главная цель этого этапа – отделить качественные описания от некачественных и исправить последние в случаях, где это возможно. Для решения этой задачи естественно использовать большую языковую модель, такую как gpt4.1-mini (OpenAI),

которая доступна по API. Эта модель была выбрана как достаточно быстрая и при этом качественная. Инструкция для модели выглядит следующим образом:

> You will be given a list of objects and their captions in form of ”id. object: caption”. Please validate captions. Good caption contain the following: - Object name (or synonym); - Material of which the object is made. If caption is good, please response ”id. Good caption.”. If you can restore missed material from the context or from the object category, update caption by adding material name to it and return ”id. new caption”. Otherwise, return ”id. Bad caption.”. Do not return any additional info.

В этой инструкции приведены два критерия, по которым необходимо оценить описание: присутствие категории объекта (или синонима) и присутствие материала. Форматы ввода и вывода также явно указаны, в частности это облегчает постобработку ответов модели. Вторая часть запроса отвечает за восстановление материала из контекста описания.

В процессе валидации удалось выделить следующие воспроизводящиеся ошибки:

- модель иногда добавляет материал в описания объектов, с которыми нельзя ассоциировать материал: например, ”an apple made of food”;
- относительно часто встречается артефакт повторения номера объекта в ответе: ”2. 2. Good caption.”;
- несмотря на прямую инструкцию, модель может возвращать рассуждения или просто воспроизводить полученное описание; это может быть также повторение некачественной надписи вместе с исправленной версией.

Перечисленные особенности усложняют постобработку ответов модели и ухудшают качество разметки.

На изображении 4 показаны результаты валидации и коррекции описаний. В области (1) красным отмечены некорректные описания, а зеленым – корректные. В области (2) показано, как выглядят доработанные описания: в обоих случаях добавлено слово ”fabric”. В области (3) показаны результаты валидации доработанных описаний. Отметим, что описание 4 для левого изображения отмечено как неправильное, хотя в нем указан материал автобуса: в данном случае не хватает категории объекта.

В таблицах 2 и 3 показаны результаты валидации: видно, что около 20% всех некорректных описаний были доразмечены именно на этом этапе.

Таблица 2: Статистика по качеству размеченных описаний для тренировочной выборки датасета MatLVIS.

Итерация разметки	Первая итерация, кол-во описаний	%	Вторая итерация (доразметка плохих описаний), кол-во описаний	%
Хорошие начальные описания	620798	48,88	757443	59,64
Плохие описания (автоматическая разметка с применением gpt4.1-mini)	649163	51,12	512518	40,36
Отредактированные описания (автоматическая разметка с применением gpt4.1-mini)	279026	21,97	499621	39,34
Описания, размеченные как хорошие на основе категории объекта	184725	14,55	0	0,00
Плохие описания, оставшиеся после пост-обработки	185412	14,60	12897	1,02
Всего	1269961		1269961	

Первая итерация автоматической разметки заканчивается валидацией, и из всех описаний примерно 15% должны быть размечены заново.

4.4 Доразметка датасета

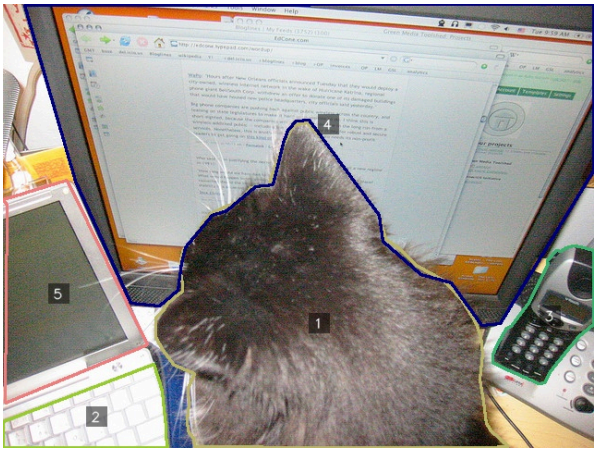
Как и основная часть разметки, доразметка произведена с применением QWEN2.5-VL. На этом этапе подход SoM уже не используется, вместо него на вход подается обрезанное с большим (около 50 пикселей) отступом изображение, на котором целевой объект обведен ярким контрастным контуром. Полученное описание снова проходит валидацию, которая описана в предыдущем разделе.

4.5 Выделение категорий материалов из описаний объектов

Для выделения материалов из описаний объектов применялась модель QWEN3 [50] (7B). Инструкция сформулирована следующим образом:

Instruction:





(1)

– Please describe the short pants labelled 1. Fit description in 10 words.
– Light-colored, casual pants worn by a person.
– Please describe the short pants labelled 2. Fit description in 10 words.
– Short pants made of fabric, worn by a person.
– Please describe the bus (vehicle) labelled 3. Fit description in 10 words.
– White and green bus parked near a grassy area.
– Please describe the bus (vehicle) labelled 4. Fit description in 10 words.
– A white bus with a green stripe and yellow parachute logo.
– Please describe the dress hat labelled 5. Fit description in 10 words.
– A black hat worn by a person.
– Please describe the parachute labelled 6. Fit description in 10 words.
– Yellow parachute on the side of a bus.
– Please describe the train (railroad vehicle) labelled 7. Fit description in 10 words.
– White train car with a black number 7.
– Please describe the taillight labelled 8. Fit description in 10 words.
– A red, plastic taillight on a white bus.

– Please describe the cat labelled 1. Fit description in 10 words.
– A grey and white cat with a long tail.
– Please describe the computer keyboard labelled 2. Fit description in 10 words.
– The keyboard is made of plastic and has white keys.
– Please describe the telephone labelled 3. Fit description in 10 words.
– A silver and black corded telephone.
– Please describe the computer monitor labelled 4. Fit description in 10 words.
– A large, black computer monitor with a white keyboard in front.
– Please describe the computer monitor labelled 5. Fit description in 10 words.
– The monitor is made of glass and metal.

⇩ ⇩

(2)

1. Light-colored, casual **fabric** pants worn by a person.
2. Short pants made of fabric, worn by a person.
3. -
4. -
5. A black **fabric** hat worn by a person.
6. -
7. -
8. A red, plastic taillight on a white bus.

1. A grey and white cat with a long tail.
2. The keyboard is made of plastic and has white keys.
3. -
4. -
5. The monitor is made of glass and metal.

⇩ ⇩

(3)

1. Light-colored, casual fabric pants worn by a person.
2. Short pants made of fabric, worn by a person.
3. The bus is made of metal, likely aluminum or steel.
4. Metal construction, likely aluminum or steel.
5. A black fabric hat worn by a person.
6. Yellow fabric canopy, likely nylon or polyester for durability.
7. White passenger train car, likely aluminum or steel.
8. A red, plastic taillight on a white bus.

1. A grey and white cat with a long tail.
2. The keyboard is made of plastic and has white keys.
3. Small, black plastic phone with a keypad and buttons.
4. Flat-screen, likely plastic or glass display with metal frame.
5. The monitor is made of glass and metal.

Рис. 4: Первый этап разметки датасета MatLVIS на примере двух изображений. (1) Разметка с применением VLM: красным показаны некорректные описания, зеленым – корректные. (2) Валидация с применением LLM: зеленым показаны изменения в сгенерированных описаниях. (3) Доразметка датасета (цветовые обозначения сохранены).

Extract the names of the **materials** mentioned in the following caption. Only return the names of physical materials such as ”leather”, ”rubber”, ”wood”, ”plastic”, ”canvas”, ”metal”, etc. If no material is mentioned, return

Таблица 3: Статистика по качеству размеченных описаний для валидационной выборки датасета MatLVIS.

Итерация	Первая итерация, кол-во описаний	%	Вторая итерация (доразметка плохих описаний), кол-во описаний	%
Хорошие начальные описания	23791	46,95	29866	58,94
Плохие описания (автоматическая разметка с применением gpt4.1-mini)	26881	53,05	20806	41,06
Отредактированные описания (автоматическая разметка с применением gpt4.1-mini)	10200	20,13	20252	39,97
Описания, размеченные как хорошие на основе категории объекта	8633	17,04	0	0,00
Плохие описания, оставшиеся после пост-обработки	8048	15,88	554	1,09
Всего	50672		50672	

”n/a”. Output format: A semicolon-separated list in lowercase, like: rubber, rubber;wood, or n/a.

Таким образом, одному объекту может соответствовать как один материал, так и несколько. В процессе разметки были замечены следующие повторяющиеся ошибки:

- Выделение характеристик объекта, которые не являются материалом: ”blue material” (когда в описании нет материала), ”similar material” (в конструкциях вида ”. . . made of something or similar material”) и другие;
- Повторяющееся возвращение одного и того же списка материалов, которые не указаны в описании.

Вторая ошибка частично исправляется ”на месте” посредством проверки вхождения указанных материалов в описание для случаев, когда выделено больше 5 материалов. После выделения материалов происходит валидация с применением gpt4.1-mini, но эта проверка не всегда позволяет исправить первую ошибку. В некоторых случа-

ях полезна ручная фильтрация на основе списков допустимых (например, ”woven”, ”plush”) и недопустимых вхождений (”blue”, ”shiny”) в имя категории.

Отметим, что не для всех описаний удалось выделить материалы. В таких случаях категория материала определялась на основе категории объекта (например, для объекта ”car” это ”metal”). В результате валидации из примерно 2500 уникальных имен материалов осталось примерно 1600.

Далее с помощью gpt4.1 было составлено соответствие между полученными уникальными категориями материалов и категориями в датасетах MINC (23 категории) и DMS (51 категория).

4.6 Результаты разметки датасета

В результате разметки датасета LVIS было собрано более миллиона описаний для различных объектов. Точные значения приведены в таблице 4. Было извлечено около 1600 уникальных материалов, что значительно превышает количество категорий в датасетах с закрытым словарем.

Таблица 4: Количество объектов, уникальных категорий и уникальных материалов в датасете.

Выборка	Объекты, ед	Уникальные категории, ед	Уникальные материалы, ед	Уникальные материалы (MINC), ед	Уникальные материалы (DMS), ед
Тренировочная	1257064	1203	1590	22	51
Валидационная	649163	803	399	21	45
Всего	1307182	1203	1612	22	51

При разметке допускается соответствие двух или более материалов одному объекту, и в отдельных случаях количество материалов для одного объекта достигает 4. В датасете есть один объект класса ”автомобиль” с описанием *”Car with metallic body, glass windows, and rubber tires; likely steel or aluminum components.”* – ему соответствует 5 разных материалов. В таблице 5 приведены значения встречаемости по количеству материалов для одного объекта.

Таблица 5: Количество материалов в одном описании.

Выборка	1 материал, число изображений	2 материала, число изображений	3 материала, число изображений	4 материала, число изображений	5 материалов, число изображений
Тренировочная	1063701	178005	15033	61	1
Валидационная	42036	7351	725	1	-

На рисунках 5, 6, 7 приведены примеры размеченных изображений из валидационной выборки датасета. Рисунок 6 показывает примеры для изображений внутри помещений: спальня, гостиная и гараж. На изображении 7 приведены уличные сцены: дорога, слоны в естественной среде обитания, яхты на берегу моря. Метки материалов обозначены фиолетовым цветом.



Рис. 5: Пример размеченного изображения из датасета MatLVIS. Стол и телефон имеют описание, в котором присутствует название материала (в данном случае пластик). Поскольку банан относится к категории fruit, в его описании не обязан фигурировать материал.

На рисунке 8 приведены диаграммы распределений по встречаемости материалов для топ-20 значений. На рисунках 9 и 10 приведены аналогичные диаграммы по всем категориям из датасетов MINC и DMS. Из этих диаграмм видно, что распределения для тренировочной и валидационной выборок достаточно сильно похожи, и в то же время заметен дисбаланс классов как среди топ-20 категорий, так и в целом по всему датасету: на первые 20 категорий приходится около 87% всех объектов в датасете.





A small, thin book made of paper on a shelf.

[paper]



The flowerpot is made of ceramic or plastic gray material.

[ceramic, plastic]



A small, decorative table lamp made of metal.

[metal]



A brown wicker basket sitting on the floor.

[wicker]





A glass vase with a white flower arrangement.

[glass, flower]



Silver metal knife with black handle, placed on white cloth.

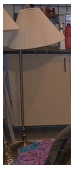
[silver]



The tablecloth is white and made of fabric.

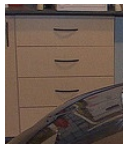
[fabric]





Metal base, fabric shade. Simple design, functional lamp.

[metal, fabric]



A metal cabinet with a wooden door and drawers.

[metal, wood]



Clear, plastic headlight with reflective inner surface.

[plastic]

Рис. 6: Примеры размеченных изображений для сцен помещений в датасете MatLVIS.





Metal SUV, likely steel or aluminum body.

[steel, aluminum]



Compact, metallic body; likely steel or aluminum.

[metal, steel, aluminum]



Black rubber wheel on a vehicle.

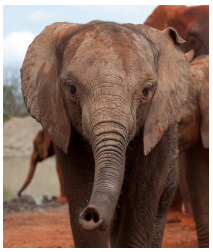
[rubber]



Rubber tire, likely made from synthetic materials.

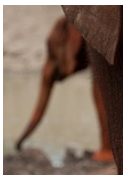
[rubber, synthetic materials]





A brown elephant with large ears and a long trunk.

[animal]



The elephant's trunk is long and made of skin.

[animal]





A dog standing in the water.

[animal]



A white cotton dress worn by a woman.

[cotton]



A white hat with a brim, likely made of straw.

[straw]



White, likely canvas or synthetic fabric; used for sailing.

[canvas, synthetic fabric]



Fabric sail with a bright white and green material.

[fabric]

Рис. 7: Примеры размеченных изображений для уличных сцен в датасете MatLVIS.

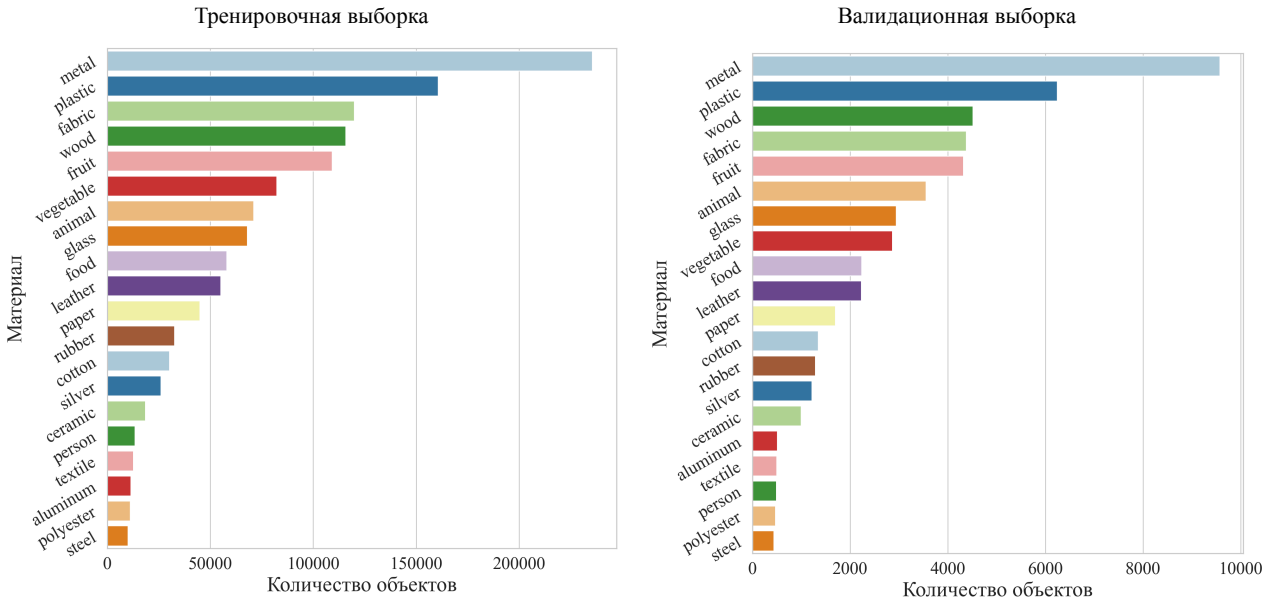


Рис. 8: Распределение топ-20 категорий материалов по количеству объектов в датасете MatLVIS. Слева приведен график для тренировочной выборки, справа – для валидационной.

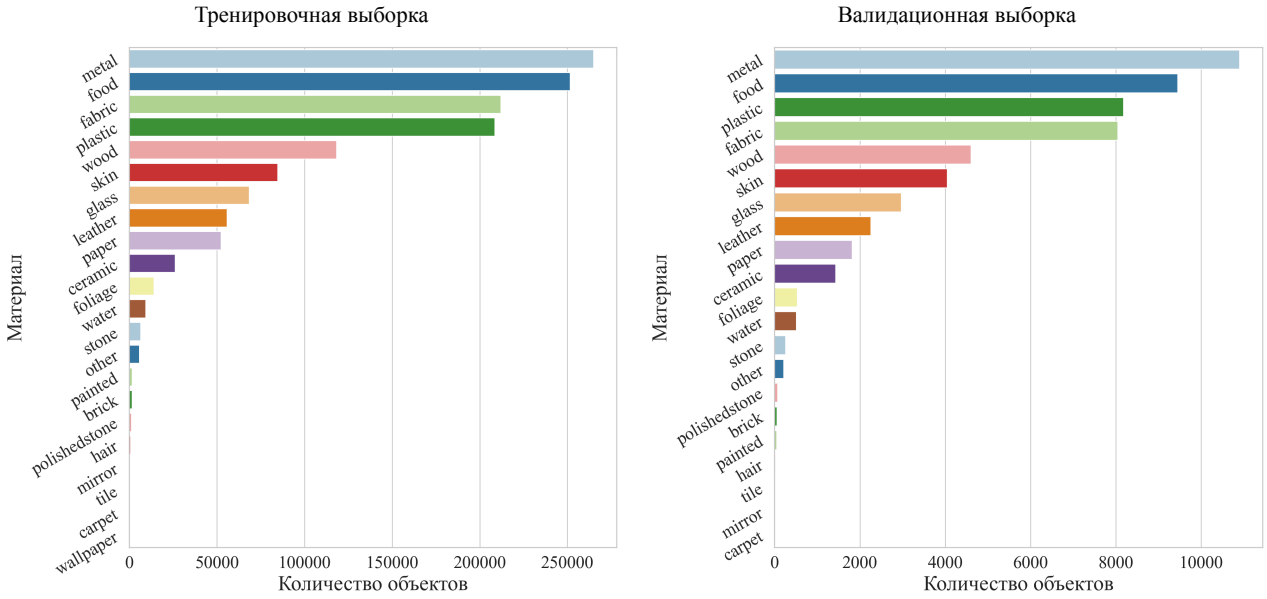


Рис. 9: Распределение категорий материалов, сопоставленных 23 категориям из датасета MINC по количеству объектов в датасете MatLVIS. Слева приведен график для тренировочной выборки, справа – для валидационной.

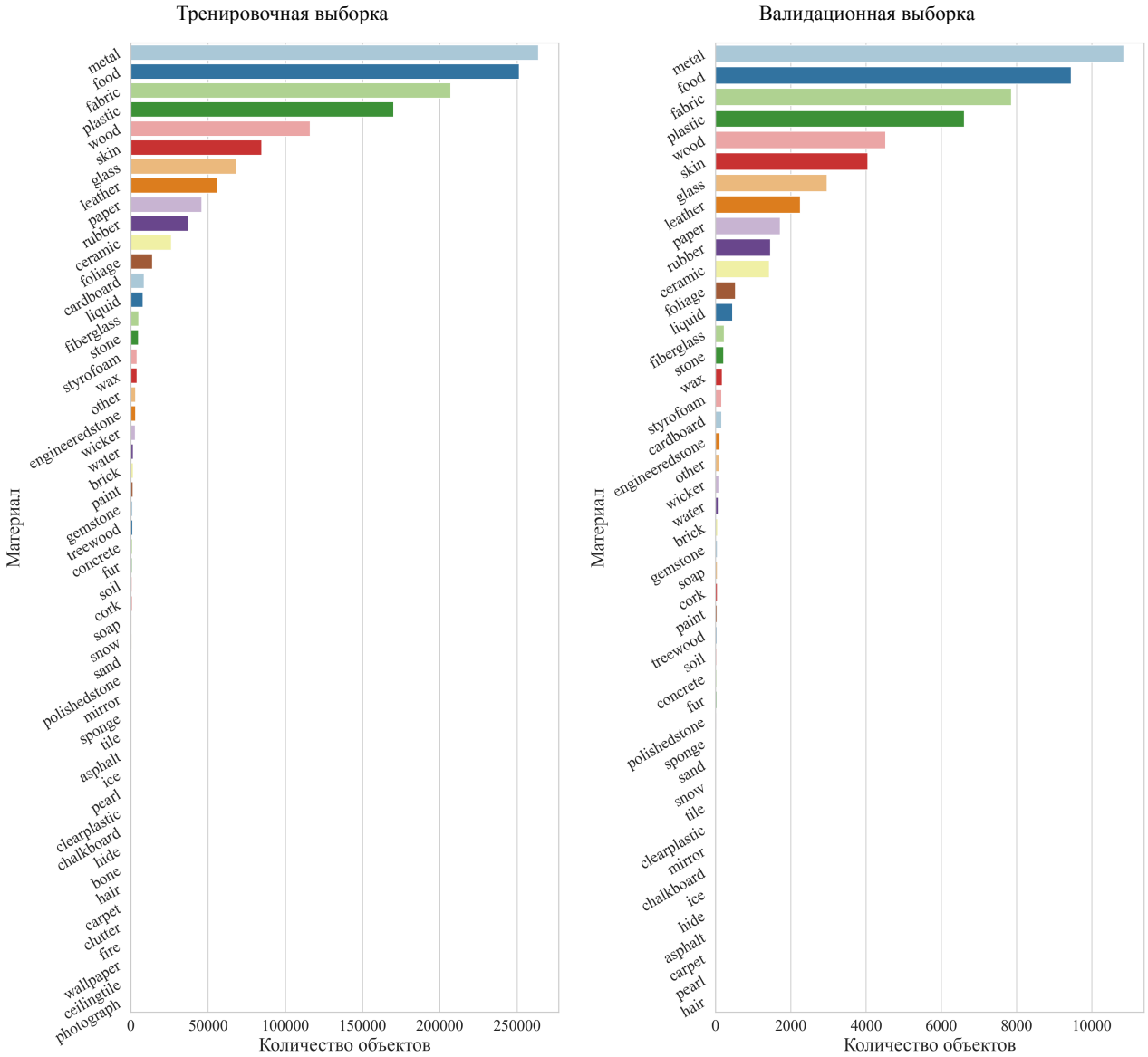


Рис. 10: Распределение категорий материалов, сопоставленных 51 категории из датасета DMS по количеству объектов в датасете MatLVIS. Слева приведен график для тренировочной выборки, справа – для валидационной.

5 Вычислительные эксперименты

В этой части работы приведены результаты экспериментов, проведенных в соответствии с рассмотренными в разделе 3 подходами: распознаванию материалов объектов на изображениях на основе статистических методов; семантическая сегментация; применение VLM; контрастивное обучение. В пятом подразделе проведен сравнительный анализ всех рассмотренных методов.

5.1 Обучение классического алгоритма на датасете MatLVIS

Обучим классический случайный лес на следующих признаках изображения объекта:

- Резкость границ, плотность границ, шероховатость (3 признака);
- Статистические признаки: эксцесс, стандартное отклонение, средняя интенсивность, энтропия, асимметрия (5 признаков);
- Признаки, извлеченные после применения фильтров Габора: стандартное отклонение, среднее значение, диапазон – для 7 фильтров (от 0° до 90° с шагом в 15°) – всего 18 признаков;
- Статистики в цветовых пространствах HSV и LAB (12 признаков);
- Гистограммы цветов – по 16 бинов на каждый из 3 каналов (48 признаков);
- LBP гистограмма – 32 бина (32 признака).

Итого получаем вектор из 118 признаков. В таблице 6 приведены указанные в разделе 3.1.3 метрики качества классификации для нескольких экспериментов. Результаты показывают, что умеренная глубина деревьев (не более 30 уровней) и относительно большое минимальное количество элементов в листе предпочтительны.

Оценка важности признаков (см. раздел 3.1, формулы (3.1), (3.2)) показывает, что наиболее полезными являются цветовые признаки LAB и HSV, а наименее полезны фильтры Габора. Сравнение важности признаков приведено в таблице 7.

Таблица 6: Метрики классификации материалов объектов в датасете MatLVIS с помощью случайного леса.

Количество деревьев	Максимальная глубина	Минимальное количество элементов в листе	P^{micro}	P^{macro}	R^{micro}	R^{macro}	F_1^{micro}	F_1^{macro}	Subset accuracy
200	30	5	0.64	**0.15**	0.22	0.04	0.32	0.05	0.22
500	30	5	0.65	0.14	0.22	0.04	0.32	0.05	0.22
500	70	5	0.69	0.13	0.20	0.03	0.32	0.04	0.21
500	30	2	**0.74**	0.13	0.17	0.02	0.27	0.03	0.18
500	30	10	0.56	0.13	**0.28**	**0.05**	**0.38**	**0.06**	**0.27**

Таблица 7: Важность различных групп признаков для решения задачи классификации.

Группа признаков	Суммарная важность	Средняя важность	Количество признаков
Цветовые признаки LAB	0.076	**0.013**	6
Цветовые признаки HSV	0.069	0.012	6
Признаки LBP	**0.346**	0.011	32
Признаки краев	0.032	0.011	3
Статистические признаки	0.033	0.007	5
Цветовая гистограмма	0.313	0.007	48
Фильтры Габора	0.112	0.006	18

5.2 Обучение моделей семантической сегментации на датасете DMS

Модели, перечисленные в разделе 3.2, были обучены на датасете DMS с помощью фреймворка MMSegmentation [22].

В таблице 8 приведены значения метрик качества сегментации для этих моделей. Для сравнения приведены также метрики модели, предоставленной авторами датасета. Эта модель имеет лучшие количественные показатели, а из обученных в данном эксперименте моделей лучшие результаты показывает HRNet.

Таблица 8: Результаты валидации сегментационных моделей в задаче определения материалов.

Модель	mIoU	mAcc
DMS	**43.23**	**59.67**
HRNet	36.55	46.71
DDRNet	33.35	43.36
Mask2Former	35.69	43.25

На изображении 11 приведены качественные результаты работы моделей на

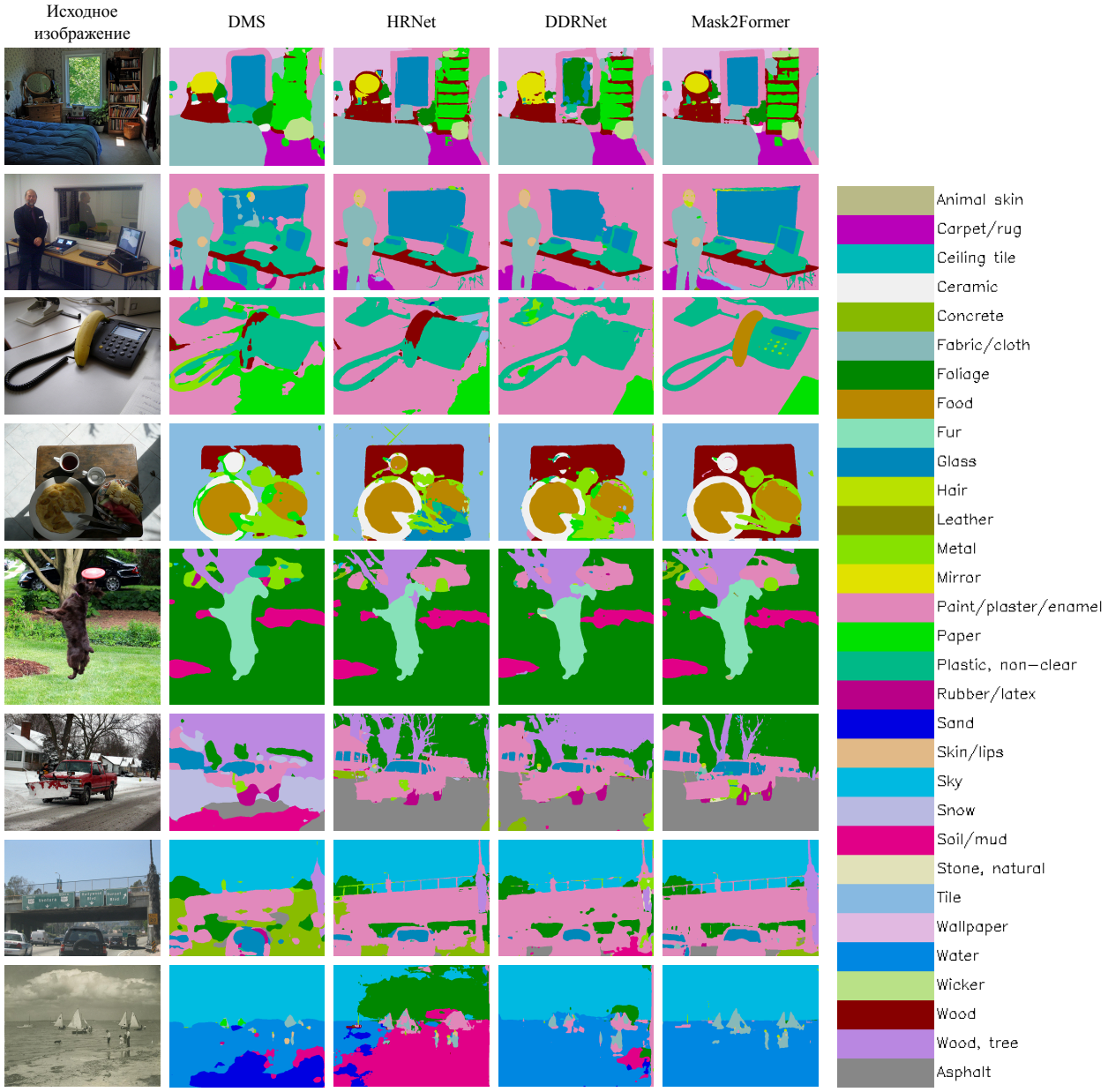


Рис. 11: Результаты работы моделей семантической сегментации на изображениях из валидационной выборки датасета MatLVIS. Справа приведена палитра.

изображениях из валидационной выборки датасета MatLVIS. Для получения данных изображений категория ”фон” была проигнорирована. Отметим, что маски, предсказанные Mask2Former, имеют наиболее гладкие границы и в некоторых случаях более корректно предсказывают категории материалов. Приведенные изображения также демонстрируют особенность разметки DMS: в этом датасете многие объекты размечены категорией ”paint”. Как правило, это объекты, поверхность которых окрашена. На изображении 11 к категории ”paint” отнесены стены, автомобили, столы, мост. Такая разметка может усложнить применение обученных на DMS сегментационных

моделей в реальных задачах.

5.3 Алгоритм определения материала объекта с помощью визуально-языковых моделей

Сравним модели QWEN2.5-VL (7B) и LLaVa1.5 (7B) в режиме VQA на валидационной выборке датасета MatLVIS. Текстовая инструкция формулируется следующим образом:

> What material is this object made of? Please only name the material. Do not answer anything except material name. If there is multiple materials, separate them with ";". Use categories from the list: asphalt; bone; brick; cardboard; carpet; ceilingtile; ceramic; chalkboard; clearplastic; clutter; concrete; cork; engineeredstone; fabric; fiberglass; fire; foliage; food; fur; gemstone; glass; hair; hide; ice; leather; liquid; metal; mirror; other; paint; paper; pearl; photograph; plastic; polishedstone; rubber; sand; skin; snow; soap; soil; sponge; stone; styrofoam; tile; treewood; wallpaper; water; wax; wicker; wood.

Это достаточно строгая инструкция, потому что она в явном виде задает формат вывода модели. В случае QWEN2.5-VL предложение "Do not answer anything except material name." можно исключить из инструкции, а вот LLaVa1.5 без этой части инструкции может генерировать ответы с объяснением выбора материала.

В инструкции указан список из 51 категории в датасете DMS: это упрощает сравнение моделей из данного раздела с другими моделями. Оценка качества производится с помощью указанных в разделе 3.1.3 метрик.

Результаты сравнения приведены в таблице 9. Видно, что LLaVa1.5 превосходит QWEN2.5-VL по F_1-мере и точности.

Таблица 9: Метрики классификации материалов в датасете MatLVIS для визуально-языковых моделей.

Модель	P^{micro}	P^{macro}	R^{micro}	R^{macro}	F_1^{micro}	F_1^{macro}	Subset accuracy
QWEN2.5-VL	0.24	0.14	**0.43**	**0.13**	0.31	0.10	0.06
LLaVa1.5	**0.38**	**0.18**	0.34	0.11	**0.36**	0.10	**0.31**

5.4 Контрастивное обучение мультимодальной архитектуры на датасете MatLVIS

Проведем серию экспериментов по дообучению CLIP на датасете MatLVIS. Будем считать базовыми экспериментами дообучение CLIP с функцией потерь InfoNCE (3.15).

Раздел 5.4.1 посвящен выбору предобученной модели. В разделе 5.4.2 рассмотрено влияние предложенных модификаций архитектуры энкодера изображений, а в 5.4.3 – влияние компоненты MultiSupCon (3.13) на результаты обучения. Зафиксируем некоторые гиперпараметры экспериментов, которые не будут изменяться далее.

- Предобработка изображений во всех экспериментах одинакова: объекты обрезаются по границам прямоугольника с добавлением отступа в 10% от большего линейного размера прямоугольника. Если итоговое изображение имеет размер менее 50 пикселей, то отступ увеличивается до достижения указанного размера.
- Аугментации изображений: повороты и отражения, изменение яркости, перевод из кодировки RGB в черно-белый режим.
- Размер батча для тренировки моделей равен 1024, а для валидации – 64. Второй параметр особенно важен, поскольку некоторые метрики зависят от него.
- Число итераций при обучении модели – 10000.

5.4.1 Выбор предобученной модели

На данный момент существует множество модификаций CLIP. Это работы, посвященные альтернативным функциям потерь, более сложным архитектурам, большим объемам тренировочных данных. Несмотря на то, что в данной работе применяется стандартная функция потерь InfoNCE и её модификация, выбор предобученной модели может существенно повлиять на результаты эксперимента.

Сравним модели CLIP, SigLIP2 [53] и TULIP [54] по метрикам согласованности между эмбеддингами изображений объектов и описаний. Проведем сравнение для версий приблизительно одного размера – ViT base. Отметим, что TULIP отличается более мелким размером патча (16 против 32). Это означает, что модель обрабатывает в качестве отдельных токенов более мелкие элементы изображения. Модели также отличаются процедурами обучения:

Таблица 10: Сравнение предобученных моделей.

Модель	MCS	MRR	Recall@1	Recall@5	Recall@10
CLIP (clip-vit-base-patch32)	**0.25**	0.61	0.48	**0.80**	**0.88**
SigLIP2 (siglip2-base-patch32-256)	0.05	0.17	0.11	0.17	0.28
TULIP (tulip-B-16-224)	0.13	0.62	0.50	0.75	0.83

- При обучении **CLIP** применяется InfoNCE на большом наборе пар «изображение–текст». Модель учится проецировать обе модальности в общее пространство, где косинусная близость отражает семантическое соответствие.

- **SigLIP2** расширяет идею CLIP, модифицируя функцию потерь для более явного разделения положительных и отрицательных пар и улучшенной стабильности обучения.

- При обучении **TULIP** (Token-level Unified Language–Image Pre-training) оптимизируется не только контрастивная функция потерь, но и reconstruction компоненты: восстановление изображения на основе его эмбеддинга и восстановление описания на основе на основе текстового эмбеддинга.

Результаты сравнения приведены в таблице 10. Из них следует, что лучше всего данным в собранном датасете соответствует обычный CLIP, поэтому далее во всех экспериментах начальные веса будут соответствовать именно этой модели.

5.4.2 Эксперименты с архитектурой модели

Этот раздел посвящен исследованию влияния архитектуры модели на качество предсказаний. В качестве базовых значений будем рассматривать метрики CLIP до обучения и метрики CLIP, обученного с использованием InfoNCE. Для удобства восприятия разделим результаты на 3 таблицы по группам метрик, указанных в разделе 3.4.2.

Из наблюдений за процессом обучения известно, что средняя косинусная близость растет только в начале обучения. После первой четверти обучения эта метрика начинает падать, в то время как остальные метрики растут. Поэтому в таблицах приведены метрики для весов, отобранных как по MCS, так и по MRR.

В таблице 11 собраны метрики, которые показывают согласованность между эмбеддингами изображений и соответствующих описаний. Видно, что дообученный

Таблица 11: Метрики согласованности между эмбеддингами изображений и соответствующих текстовых описаний для разных архитектур модели.

Модель	Кодировщик глобального контекста	Функция потерь	Отбор по метрике	MCS	MRR	Recall@1	Recall@5	Recall@10
CLIP	-	-	-	0.25	0.61	0.48	0.80	0.88
CLIP	-	InfoNCE	MRR	0.43	0.85	0.75	0.97	0.98
			MCS	0.45	0.75	0.64	0.95	0.97
CLIP + fusion (MLP)	DINOv2-base	InfoNCE	MRR	0.51	0.89	0.82	0.98	0.99
			MCS	0.55	0.87	0.78	0.98	0.99
CLIP + fusion (MLP)	RADIOv2.5-base	InfoNCE	MRR	0.51	**0.91**	**0.85**	**0.99**	**1.00**
			MCS	0.56	0.89	0.81	0.98	0.99
CLIP + fusion (MHA)	RADIOv2.5-base	InfoNCE	MRR	0.67	0.78	0.67	0.92	0.96
			MCS	**0.71**	0.74	0.61	0.90	0.95

CLIP превосходит CLIP перед обучением, а модифицированные версии CLIP превосходят обычную версию модели. Лучшие результаты по MRR и Recall показывает архитектура, в которой эмбеддинги CLIP объединяются с эмбеддингами RADIOv2.5-base с помощью MLP. Архитектура, в которой для того же применяется MHA, показывает результаты на уровне обычного CLIP по тем же метрикам, однако значительно превосходит другие модели по MCS. Помимо улучшения метрик, добавление контекстной информации позитивно влияет на стабильность обучения.

Достаточно высокие значения MRR и Recall показывают, что CLIP обучается корректно сопоставлять изображение и описание, причем дополнительная информация о контексте действительно улучшает качество работы модели. Однако этих метрик недостаточно, чтобы говорить о том, что модель выучивает материал объекта. Оценить последнее помогает вторая часть валидации модели: оценка эмбеддингов изображений (таблица 12).

Лучшие результаты показывают модели, отобранные по MCS, а также модифицированная архитектура CLIP с объединением признаков на основе механизма внимания. При этом ”внутренние” метрики для эмбеддингов достаточно невысокие: коэффициент Пирсона около 0.3-0.4 показывает слабую скоррелированность между сходством эмбеддингов и сходством категорий материалов (что, конечно, лучше отсутствия корреляции). Silhouette score около нуля показывает пересечение кластеров, а ARI около 0.2 лучше, чем произвольное назначение класса материала: в таком случае метрика принимала бы значение, близкое к нулю. F_1 и mAP принимают приемлемые значения. Стоит отметить, что валидация происходит с опорой на все уникальные категории материалов, то есть около 1600 классов. Учитывая дисбаланс классов и их достаточно большое количество, а также наличие синонимов, наблюдаемые метрики можно рассматривать как приемлемые. Интересно, что луч-

Таблица 12: Метрики качества эмбеддингов изображений для разных архитектур модели.

Модель	Кодировщик глобального контекста	Функция потерь	Отбор по метрике	Pearson corr^m	Silhouette score^m	ARI^m	F_1^m	mAP^m
CLIP	-	-	-	0.23	**-0.02**	0.17	0.80	0.30
CLIP	-	InfoNCE	MRR	0.34	**-0.02**	0.17	0.77	0.32
			MCS	**0.36**	**-0.02**	**0.18**	0.78	**0.36**
CLIP + fusion (MLP)	DINOv2-base	InfoNCE	MRR	0.27	-0.03	0.16	0.79	0.29
			MCS	0.30	-0.04	0.14	0.79	0.32
CLIP + fusion (MLP)	RADIOv2.5-base	InfoNCE	MRR	0.27	-0.03	0.15	0.79	0.29
			MCS	0.32	-0.03	0.15	0.79	0.33
CLIP + fusion (MHA)	RADIOv2.5-base	InfoNCE	MRR	0.34	-0.07	0.17	**0.81**	0.32
			MCS	**0.36**	-0.07	0.17	**0.81**	0.34

Таблица 13: Метрики классификации материалов на основе эмбеддингов изображений для разных архитектур модели.

Модель	Кодировщик глобального контекста	Функция потерь	Отбор по метрике	P^{micro}	P^{macro}	R^{micro}	R^{macro}	F_1^{micro}	F_1^{macro}	Subset accuracy
CLIP	-	-	-	0.47	0.22	0.78	0.23	0.59	0.21	0.37
CLIP	-	InfoNCE	MRR	0.56	0.25	0.84	0.30	0.68	0.26	0.49
			MCS	**0.60**	0.26	0.78	0.25	0.68	0.24	0.51
CLIP + fusion (MLP)	DINOv2-base	InfoNCE	MRR	0.58	0.26	0.84	0.30	0.69	0.26	0.50
			MCS	0.56	0.25	0.84	0.28	0.67	0.25	0.47
CLIP + fusion (MLP)	RADIOv2.5-base	InfoNCE	MRR	**0.60**	**0.29**	0.86	0.33	**0.70**	**0.29**	**0.52**
			MCS	0.51	0.27	**0.89**	**0.35**	0.65	0.27	0.42
CLIP + fusion (MHA)	RADIOv2.5-base	InfoNCE	MRR	0.54	0.22	0.82	0.26	0.65	0.23	0.45
			MCS	0.52	0.20	0.81	0.24	0.63	0.20	0.43

шие метрики в этом эксперименте находятся ”в оппозиции” с предыдущим набором метрик (таблица 11).

Наконец, рассмотрим применимость обученных картиночных эмбеддингов для решения downstream задач: обучим на них классификатор для распознавания категорий материалов. Воспользуемся набором категорий для датасета DMS.

В качестве классификатора воспользуемся двухслойным MLP. Поскольку решается задача классификации с несколькими метками, традиционную выходную функцию активации softmax заменим на сигмоиду, в качестве функции потерь воспользуемся бинарной кросс-энтропией 2.1.

Метрики в таблице 13 показывают, что лучшие эмбеддинги строит модифицированный CLIP с объединением признаков на основе MLP. Это согласуется с результатами оценки согласованности текстовых и картиночных эмбеддингов (таблица 11).

Отметим, что обучение обычного CLIP дает наибольший прирост качества в таких категориях, как *leather, ceramic, foliage*. Добавление информации о контексте в эксперименте CLIP + fusion (MLP) (с контекстными эмбеддингами RADIOv2.5) сильнее всего улучшает качество классификации категорий *ceramic, wood, paper*. На

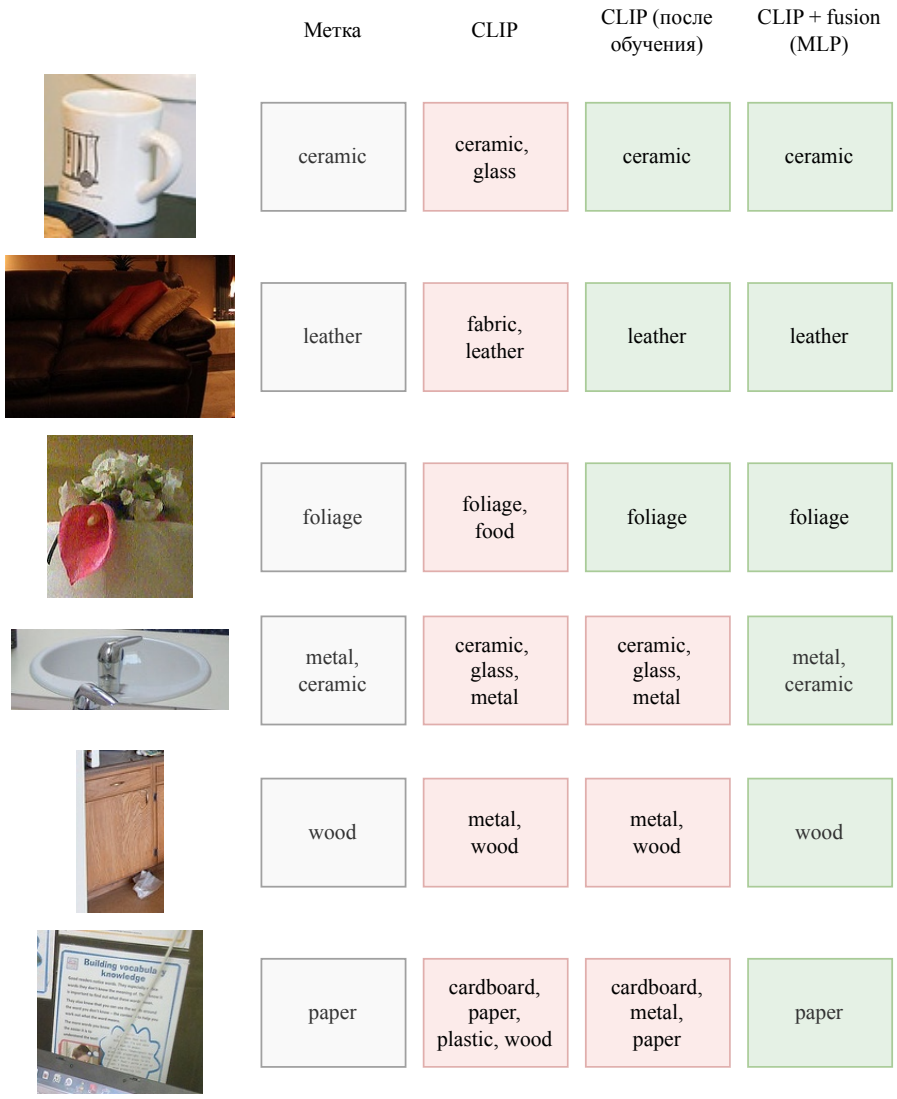


Рис. 12: Примеры работы классификаторов на основе предобученного CLIP, обученного CLIP и обученного CLIP + fusion (MLP).

изображении 12 приведены примеры работы моделей для изображений объектов с материалами из перечисленных категорий.

5.4.3 Эксперименты с функцией потерь

Этот раздел работы посвящен модифицированной функции потерь, в которой помимо InfoNCE (3.15) присутствует компонента MultiSupCon (3.13) для категорий объектов и материалов. Как было отмечено в секции 3.4.1, эта функция потерь служит для сближения эмбеддингов объектов с синонимичными категориями и отдаления эмбеддингов объектов с категориями-антонимами.

Проведены следующие эксперименты:

- **Fine-tuning текстовой части CLIP с MultiSupCon.** Поскольку в текстах явно указаны категории и материалы, обучение текстовых эмбеддингов должно пройти быстро и стабильно. Таким образом текстовый модуль предварительно ”адаптируется” с помощью MultiSupCon до начала совместного обучения, чтобы улучшить качество начальных представлений и ускорить сходимость основной задачи.

- **Совместное применение MultiSupCon к изображениям и текстам.** Контрастивная MultiSupCon применяется одновременно к эмбеддингам обеих модальностей, стимулируя более плотное выравнивание семантически сходных пар и разрежение негативных пар в объединённом пространстве.

- **MultiSupCon только для изображений объектов.** Основная цель — получение качественных визуальных представлений. Обучение текстового модуля происходит без изменений, в то время как за счет MultiSupCon фокус сделан на изображениях. Это должно максимизировать внутриклассовое сходство визуальных эмбеддингов.

- **MultiSupCon только для описаний объектов.** Текстовую модель легче обучить для оптимизации MultiSupCon, поэтому ограничимся обучением эмбеддингов описаний. Визуальный модуль при этом обучается только через InfoNCE, получая косвенный эффект от улучшенных текстовых представлений.

В таблице 14 показаны метрики согласованности между эмбеддингами изображений и описаний. Видно, что лучшие метрики показывают модели, которые имеют модифицированную архитектуру.

Достаточно высокие метрики показывает эксперимент, в котором текстовую часть CLIP предварительно обучили для минимизации функции MultiSupCon (3.13): в этом эксперименте основное обучение проводилось с функцией потерь InfoNCE 3.15.

Интересно, что в случае обучения CLIP с комбинированной функцией потерь метрики почти не отличаются от метрик для необученного CLIP. Можно предположить, что оптимизация метрики MultiSupCon препятствует смещению весов модели от их начального положения: в пользу этой гипотезы свидетельствует график функ-

Таблица 14: Метрики согласованности между эмбеддингами изображений и соответствующих текстовых описаний для разных архитектур модели.

Модель	Функция потерь	Отбор по метрике	MCS	MRR	Recall@1	Recall@5	Recall@10
CLIP	-	-	0.25	0.61	0.48	0.80	0.88
CLIP	InfoNCE	MRR	0.43	0.85	0.75	0.97	0.98
		MCS	0.45	0.75	0.64	0.95	0.97
CLIP (текстовый энкодер предобучен с MultiSupCon)	InfoNCE	MRR	0.43	0.85	0.77	0.95	0.98
		MCS	0.47	0.83	0.72	0.95	0.98
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.24	0.73	0.59	0.91	0.94
		MCS	0.24	0.74	0.61	0.92	0.94
CLIP	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.25	0.72	0.58	0.92	0.94
		MCS	0.25	0.72	0.58	0.92	0.94
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	0.17	0.71	0.58	0.88	0.97
		MCS	0.17	0.71	0.58	0.88	0.97
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.49	**0.91**	**0.84**	**0.99**	**1.00**
		MCS	0.52	0.89	0.81	0.98	0.99
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.51	**0.91**	**0.84**	**0.99**	**1.00**
		MCS	**0.54**	0.89	0.81	0.98	0.99
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	0.46	**0.91**	**0.84**	**0.99**	**1.00**
		MCS	0.50	0.89	0.81	0.98	0.99
CLIP + fusion (MHA)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.37	0.78	0.67	0.93	0.96
		MCS	0.37	0.75	0.62	0.91	0.95

ции потерь, который имеет достаточно нестабильное поведение и слабо падает по сравнению с базовым экспериментом.

В таблице 15 приведены метрики качества эмбеддингов: они мало отличаются от результатов для экспериментов из предыдущего раздела.

В таблице 16 приведены метрики классификации материалов объектов для обученного на картиночных эмбеддингах классификатора. Архитектура модели, как и в предыдущем разделе, представляет двухслойный MLP.

Следует отметить, что лучшие показатели метрик продемонстрировала конфигурация модели, в которой MultiSupCon применялась только к текстовым эмбеддингам. Это объясняется тем, что оптимизация текстового энкодера с использованием MultiSupCon проходит более эффективно. Напомним, что данная функция потерь нацелена на сближение эмбеддингов объектов, изготовленных из сходных материалов, и на разрежение эмбеддингов остальных объектов. Поскольку информация о материале явно представлена в текстовом описании, текстовый энкодер способен надёжно выявлять соответствующие признаки. Кроме того, более информативные текстовые эмбеддинги дают дополнительную поддержку при оптимизации InfoNCE, которая связывает текстовые и визуальные представления.

Эксперименты, проведенные в этом разделе, показывают, что применение

Таблица 15: Метрики качества эмбеддингов изображений для разных архитектур модели.

Модель	Функция потерь	Отбор по метрике	Pearson corr^m	Silhouette score^m	ARI^m	F_1^m	mAP^m
CLIP	-	-	0.23	-0.02	0.17	0.80	0.30
CLIP	InfoNCE	MRR	0.34	-0.02	0.17	0.77	0.32
		MCS	**0.36**	-0.02	0.18	0.78	**0.36**
CLIP (текстовый энкодер предобучен с MultiSupCon)	InfoNCE	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.26	-0.01	**0.19**	**0.83**	0.32
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.26	-0.01	**0.19**	**0.83**	0.32
CLIP	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.26	-0.01	**0.19**	**0.83**	0.32
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.26	-0.01	**0.19**	**0.83**	0.32
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.25	-0.01	0.18	0.82	0.32
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.25	-0.01	0.18	0.82	0.32
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	0.25	-0.01	0.18	**0.83**	0.32
		MCS	0.25	-0.01	0.18	0.82	0.32
CLIP + fusion (MHA)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	**0.36**	-0.01	0.16	0.81	0.35
		MCS	**0.36**	-0.01	0.17	0.82	0.35

Таблица 16: Метрики классификации материалов объектов на основе эмбеддингов изображений для разных архитектур модели.

Модель	Функция потерь	Отбор по метрике	P^{micro}	P^{macro}	R^{micro}	R^{macro}	$\text{F}_1^{\text{micro}}$	$\text{F}_1^{\text{macro}}$	Subset accuracy
CLIP	-	-	0.47	0.22	0.78	0.23	0.59	0.21	0.37
CLIP	InfoNCE	MRR	0.56	0.25	0.84	0.30	0.68	0.26	0.49
		MCS	0.60	0.26	0.78	0.25	0.68	0.24	0.51
CLIP (текстовый энкодер предобучен с MultiSupCon)	InfoNCE	MRR	0.67	0.30	0.78	0.26	0.72	0.26	0.59
		MCS	0.60	0.28	0.78	0.25	0.68	0.25	0.52
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.48	0.24	0.78	0.23	0.59	0.21	0.37
		MCS	0.48	0.24	0.78	0.24	0.59	0.21	0.38
CLIP	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.48	0.23	0.78	0.23	0.59	0.21	0.37
		MCS	0.55	0.26	0.73	0.21	0.62	0.21	0.45
CLIP	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	0.48	0.24	0.78	0.24	0.59	0.21	0.37
		MCS	0.40	0.20	0.84	0.28	0.54	0.21	0.29
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.66	0.27	0.83	0.28	0.73	0.26	0.58
		MCS	0.59	0.24	**0.86**	0.29	0.70	0.25	0.51
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{img}}$	MRR	0.60	0.27	**0.86**	**0.31**	0.71	0.27	0.53
		MCS	0.58	0.25	**0.86**	0.29	0.69	0.25	0.50
CLIP + fusion (MLP)	InfoNCE + $\text{MSupCon}^{\text{txt}}$	MRR	**0.70**	**0.32**	0.80	0.27	**0.75**	**0.28**	**0.62**
		MCS	0.59	0.26	**0.86**	0.30	0.70	0.27	0.51
CLIP + fusion (MHA)	InfoNCE + $\text{MSupCon}^{\text{txt}}$ + $\text{MSupCon}^{\text{img}}$	MRR	0.55	0.22	0.83	0.26	0.66	0.23	0.47
		MCS	0.60	0.23	0.78	0.23	0.68	0.22	0.52

MultiSupCon 3.13 вместе с InfoNCE 3.15 улучшает качество эмбеддингов CLIP и помогает в последующей классификации, однако ключевым аспектом успешного обучения является учет контекста.

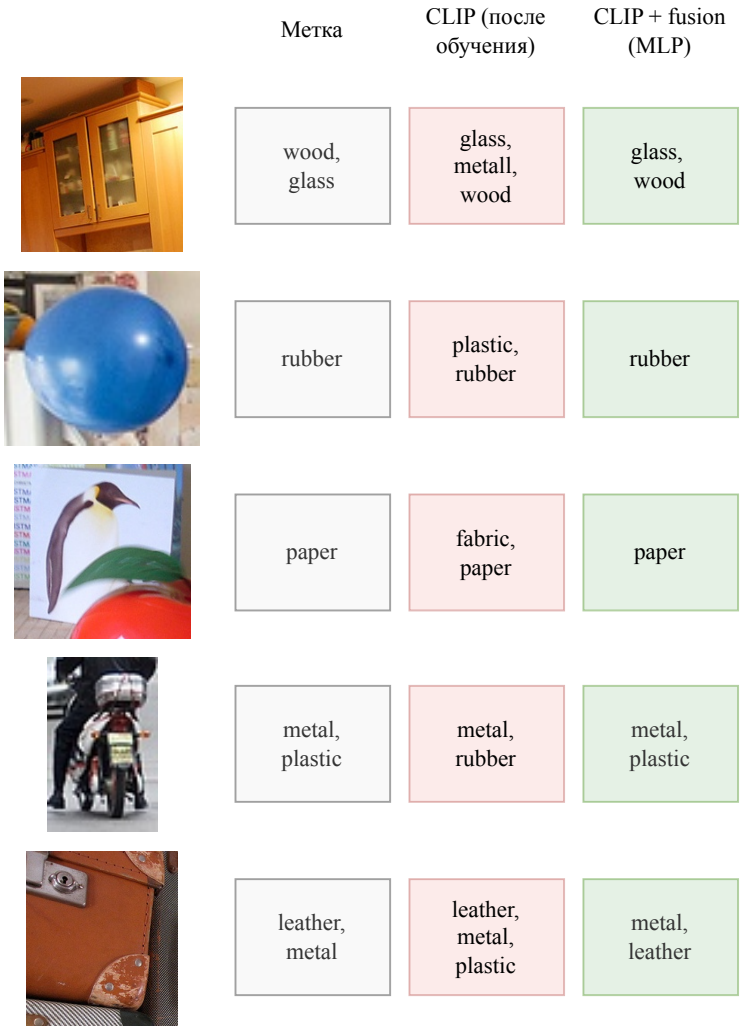


Рис. 13: Примеры работы классификаторов на основе предобученного CLIP, обученного CLIP и обученного с применением модифицированной функци потерь CLIP + fusion (MLP).

Модель с модифицированной архитектурой, обученная на InfoNCE+MultiSupCon^{txt}, дает наибольший прирост качества по сравнению с обученным CLIP в категориях *wood, rubber, paper, plastic, leather*. На изображении 13 приведены примеры работы указанных моделей на изображениях объектов из перечисленных категорий.

5.5 Сравнение различных методов распознавания материалов на датасете MatLVIS

В данном разделе проведено сравнение различных подходов по распознаванию материалов на валидационной выборке датасета MatLVIS. Как и прежде, решается задача классификации с несколькими метками. Для оценки моделей применяются метрики классификации из раздела 3.1.3.

Расширим набор рассмотренных ранее методов за счет следующих подходов:

- Инференс модели, обученной на DMS (в данном случае это модель, предложенная авторами статьи);
- Обучение ResNet50 для классификации с несколькими метками;
- Инференс YOLOE, CLIP, SigLIP2 и TULIP в режиме zero-shot классификации;

Рассмотрим подробнее каждый из них.

Инференс модели DMS. Модель, предложенная в работе DMS, решает задачу семантической сегментации. Для перевода предсказаний в модели в классы применяется бинарная маска объекта и фильтрация по площади: считаем, что объект состоит из материала, если площадь соответствующей части маскированной области превышает 25% от площади маски.

Обучение ResNet50. ResNet – это классическая сверточная архитектура, которая применяется как в качестве самостоятельной модели для решения различных задач компьютерного зрения, так и в составе более сложных архитектур. Обучать модель будем аналогично подходу, предложенному для обучения классификатора в разделе 5.4.2.

Инференс YOLOE, CLIP, SigLIP2 и TULIP. Рассмотрим следующий алгоритм классификации с применением этих моделей следующий:

- Построение векторных представления для набора категорий из датасета DMS с помощью текстовых энкодеров перечисленных моделей.
- Применение визуальных энкодеров моделей для формирования векторного представления изображения объекта.
- Поиск эмбеддинга категории, наиболее близкого к эмбеддингу изображения по косинусной близости.

Текстовые энкодеры и энкодеры изображений CLIP, SigLIP2 и TULIP строят отображения в одном пространстве признаков. YOLOE, в отличие от этих моделей, также предсказывает ограничивающий прямоугольник и маску объекта.

В таблице 17 приведены результаты сравнения. Заметим, что метрики, полученные макро-усреднением, невысоки (не более 0.3 для F_1-меры). Поскольку обучение не было оптимизировано для задачи с дисбалансом классов, такое поведение вполне естественно.

В остальном, худшие результаты показывают мультимодальные подходы, в основе которых лежит контрастивное обучение (CLIP, SigLIP2, TULIP) и YOLOE. Это можно объяснить спецификой данных, на которых обучались перечисленные модели: поскольку обычные описания объектов редко сфокусированы на материалах, в обучающих выборках не набралось достаточно примеров для решения задачи классификации материалов. Немного лучше справляются LLaVa1.5 и DMS: это методы, которые не обучались на датасете MatLVIS. Далее идет случайный лес на основе статистических признаков: эта модель традиционно слабее более сложной сверточной архитектуры ResNet, которая способна выделять более сложные и не предопределенные признаки. Наконец, лучшие результаты показывает простой классификатор над признаками усовершенствованной модели на основе CLIP: 70% предсказаний модели верны и в 80% случаев она корректно распознает по крайней мере один материал объекта. При этом в 62% случаев совпадение по меткам материалов полное.

На изображении 14 приведены примеры работы моделей.

Таблица 17: Метрики классификации материалов в датасете MatLVIS для различных методов.

Модель	Датасет для обучения	P^{micro}	P^{macro}	R^{micro}	R^{macro}	F_1^{micro}	F_1^{macro}	Subset accuracy
Random Forest	MatLVIS	<u>0.56</u>	0.13	0.28	0.05	<u>0.38</u>	0.06	0.27
ResNet50	MatLVIS	0.53	0.22	**0.81**	0.26	0.64	<u>0.22</u>	<u>0.45</u>
DMS	DMS	0.38	0.13	0.37	0.21	0.37	0.10	0.35
LLaVa1.5	-	0.38	<u>0.18</u>	0.34	0.11	0.36	0.10	0.31
YOLOE	-	0.17	0.06	0.002	0.01	0.004	0.004	0.001
CLIP	-	0.04	0.04	0.44	**0.41**	0.07	0.05	0.01
SigLIP2	-	0.02	0.02	0.42	<u>0.39</u>	0.04	0.03	0.0
TULIP	-	0.08	0.07	0.38	0.36	0.13	0.07	0.04
CLIP + Fusion (MLP) (ours)	MatLVIS	**0.70**	**0.32**	<u>0.80</u>	0.27	**0.75**	**0.28**	**0.62**



Рис. 14: Примеры работы различных моделей для решения задачи классификации материалов объектов.

6 Заключение

В настоящей работе подробно исследована задача распознавания материалов объектов в трёхмерных сценах по их RGB-изображениям. Результаты работы сводятся к следующим пунктам:

- **Создан датасет MatLVIS.** Разработано расширение датасета LVIS, включающее корпус из более 1.2 млн естественных описаний объектов с явным указанием материалов. Датасет содержит около 1600 уникальных категорий материалов, что значительно превышает возможности существующих датасетов с закрытым словарем (не более 60 категорий). Предложен алгоритм автоматической разметки с применением визуально-языковых и больших языковых моделей, обеспечивающий высокое качество аннотаций.

- **Проанализированы классические и современные методы.** Проведено комплексное сравнение подходов на основе статистических признаков изображений, семантической сегментации, и инференса визуально-языковых моделей. Показано, что традиционные мультимодальные модели (CLIP, SigLIP2, TULIP) демонстрируют низкое качество распознавания материалов (F1-мера не превышает 0.13) из-за недостаточной представленности информации о материалах в обучающих данных.

- **Разработан контрастивный мультимодальный метод.** Предложена модифицированная архитектура на основе CLIP, объединяющая контрастивную функцию потерь InfoNCE с расширенной функцией MultiSupCon для учета сходства материалов и семантических классов объектов. Введен модуль объединения контекстной информации, улучшающий качество эмбеддингов изображений. Линейный классификатор, обученный на полученных эмбеддингах, достиг точности 70% и полного совпадения меток в 62% случаев, превзойдя все рассмотренные альтернативы.

Экспериментальные результаты подтверждают эффективность предложенного подхода для решения задачи многометочной классификации материалов объектов. Метод показывает устойчивую работу на разнообразных категориях материалов и превосходит как классические алгоритмы машинного обучения, так и современные модели глубокого обучения.

Дальнейшие исследования могут быть направлены на разработку комплексного решения для оценки физических свойств объектов путем интеграции предложенного метода распознавания материалов с базами данных физических характеристик материалов. Перспективным направлением является также расширение подхода на другие модальности (тактильные данные, спектральная информация) для повышения точности определения свойств материалов.

Данная работа вносит вклад в методологию мультимодального контрастивного обучения для задач распознавания материалов. Созданный датасет и эффективная модель имеют практическое значение для приложений компьютерного зрения, особенно в области робототехники, где требуется физическое взаимодействие с объектами.

Список литературы

1. *Gabor D.* Theory of Communication // Journal of the Institution of Electrical Engineers – Part III: Radio and Communication Engineering. — 1946. — Т. 93, № 26. — С. 429—457.
2. *Sobel I.*, *Feldman G.* A 3×3 Isotropic Gradient Operator for Image Processing : Technical Report / Stanford Artificial Intelligence Project (SAIL). — Stanford, CA, 1968.
3. *Breiman L.* Random Forests // Machine Learning. — 2001. — Т. 45, № 1. — С. 5—32. — DOI: 10.1023/A:1010933404324.
4. *Lazebnik S.*, *Schmid C.*, *Ponce J.* A sparse texture representation using local affine regions // IEEE transactions on pattern analysis and machine intelligence. — 2005. — Т. 27, № 8. — С. 1265—1278.
5. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation / A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder [и др.] // APL materials. — 2013. — Т. 1, № 1.
6. OpenSurfaces: A richly annotated catalog of surface appearance / S. Bell, P. Upchurch, N. Snavely, K. Bala // ACM Transactions on graphics (TOG). — 2013. — Т. 32, № 4. — С. 1—17.
7. Describing textures in the wild / M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2014. — С. 3606—3613.
8. *Sharan L.*, *Rosenholtz R.*, *Adelson E. H.* Accuracy and speed of material categorization in real-world images // Journal of Vision. — 2014. — Т. 14, № 10.
9. *Cimpoi M.*, *Maji S.*, *Vedaldi A.* Deep filter banks for texture recognition and segmentation // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — С. 3828—3836.
10. Material Recognition in the Wild with the Materials in Context Database / S. Bell, P. Upchurch, N. Snavely, K. Bala // Computer Vision and Pattern Recognition (CVPR). — 2015.

11. Material recognition in the wild with the materials in context database / S. Bell, P. Upchurch, N. Snavely, K. Bala // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2015. — С. 3479—3487.

12. Deep learning for tactile understanding from visual and haptic data / Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, T. Darrell // 2016 IEEE international conference on robotics and automation (ICRA). — IEEE. 2016. — С. 536—543.

13. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin // Advances in neural information processing systems. — 2017. — Т. 30.

14. Differential angular imaging for material recognition / J. Xue, H. Zhang, K. Dana, K. Nishino // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. — 2017. — С. 764—773.

15. *Oord A. v. d.*, *Li Y.*, *Vinyals O.* Representation learning with contrastive predictive coding // arXiv preprint arXiv:1807.03748. — 2018.

16. 3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera / I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, S. Savarese // Proceedings of the IEEE International Conference on Computer Vision. — 2019.

17. *Gupta A.*, *Dollar P.*, *Girshick R.* Lvis: A dataset for large vocabulary instance segmentation // Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. — 2019. — С. 5356—5364.

18. A simple framework for contrastive learning of visual representations / T. Chen, S. Kornblith, M. Norouzi, G. Hinton // International conference on machine learning. — PmLR. 2020. — С. 1597—1607.

19. An image is worth 16x16 words: Transformers for image recognition at scale / A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly [и др.] // arXiv preprint arXiv:2010.11929. — 2020.

20. Deep high-resolution representation learning for visual recognition / J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang [и др.] // IEEE transactions on pattern analysis and machine intelligence. — 2020. — Т. 43, № 10. — С. 3349—3364.

21. Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions / J. Wald, H. Dhamo, N. Navab, F. Tombari // Conference on Computer Vision and Pattern Recognition (CVPR). — 2020.

22. *MMSegmentation Contributors*. OpenMMLab Semantic Segmentation Toolbox and Benchmark. — 07.2020. — URL: `https://github.com/open-mmlab/mmsegmentation`.

23. *Schwartz G.*, *Nishino K.* Recognizing Material Properties from Images // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2020. — Т. 42, № 8. — С. 1981—1995. — DOI: `10.1109/TPAMI.2019.2907850`.

24. Supervised contrastive learning / P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan // Advances in neural information processing systems. — 2020. — Т. 33. — С. 18661—18673.

25. *Trémeau A.*, *Xu S.*, *Muselet D.* Deep learning for material recognition: most recent advances and open challenges // arXiv preprint arXiv:2012.07495. — 2020.

26. *Yuan Y.*, *Chen X.*, *Wang J.* Object-Contextual Representations for Semantic Segmentation. — 2020.

27. 3d-future: 3d furniture shape with texture / H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. Maybank, D. Tao // International Journal of Computer Vision. — 2021. — С. 1—25.

28. *Cheng B.*, *Schwing A. G.*, *Kirillov A.* Per-Pixel Classification is Not All You Need for Semantic Segmentation //. — 2021.

29. Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes / Y. Hong, H. Pan, W. Sun, Y. Jia // arXiv preprint arXiv:2101.06085. — 2021.

30. Learning transferable visual models from natural language supervision / A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark [и др.] // International conference on machine learning. — PmLR. 2021. — С. 8748—8763.

31. *Roberts M.*, *Ramapuram J.*, *Ranjan A.*, *Kumar A.*, *Bautista M. A.*, *Paczan N.*, *Webb R.*, *Susskind J. M.* Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding // International Conference on Computer Vision (ICCV) 2021. — 2021.

32. Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes / H. Pan, Y. Hong, W. Sun, Y. Jia // IEEE Transactions on Intelligent Transportation Systems. — 2022.

33. 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.

34. MINERVAS: Massive INterior EnviRonments VirtuAl Synthesis / H. Ren, H. Zhang, J. Zheng, J. Zheng, R. Tang, Y. Huo, H. Bao, R. Wang. — 2022. — arXiv: 2107.06149 [cs.CV]. — URL: https://arxiv.org/abs/2107.06149.

35. Multimodal material segmentation / Y. Liang, R. Wakaki, S. Nobuhara, K. Nishino // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2022. — С. 19800—19808.

36. *Upchurch P.*, *Niu R.* A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing // European Conference on Computer Vision (ECCV). — 2022.

37. *Upchurch P.*, *Niu R.* A dense material segmentation dataset for indoor and outdoor scene parsing // European conference on computer vision. — Springer. 2022. — С. 450—466.

38. Dinov2: Learning robust visual features without supervision / M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby [и др.] // arXiv preprint arXiv:2304.07193. — 2023.

39. Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V / J. Yang, H. Zhang, F. Li, X. Zou, C. Li, J. Gao. — 2023. — arXiv: 2310.11441 [cs.CV]. — URL: https://arxiv.org/abs/2310.11441.

40. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v / J. Yang, H. Zhang, F. Li, X. Zou, C. Li, J. Gao // arXiv preprint arXiv:2310.11441. — 2023.

41. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality / W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing. — 03.2023. — URL: https://lmsys.org/blog/2023-03-30-vicuna/.

42. Improved Baselines with Visual Instruction Tuning / H. Liu, C. Li, Y. Li, Y. J. Lee. — 2024. — arXiv: 2310.03744 [cs.CV]. — URL: https://arxiv.org/abs/2310.03744.

43. Improved baselines with visual instruction tuning / H. Liu, C. Li, Y. Li, Y. J. Lee // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2024. — С. 26296—26306.

44. Materobot: Material recognition in wearable robotics for people with visual impairments / J. Zheng, J. Zhang, K. Yang, K. Peng, R. Stiefelhagen // 2024 IEEE International Conference on Robotics and Automation (ICRA). — IEEE. 2024. — С. 2303—2309.

45. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution / P. Wang [и др.]. — 2024. — arXiv: 2409.12191 [cs.CV]. — URL: https://arxiv.org/abs/2409.12191.

46. *Sun Y.*, *Gu Z.*, *Yang S. B.* Probing vision and language models for construction waste material recognition // Automation in Construction. — 2024. — Т. 166. — С. 105629.

47. *Chen Y.*, *Son H.*, *Kusari A.* MatPredict: a dataset and benchmark for learning material properties of diverse indoor objects // arXiv preprint arXiv:2505.13201. — 2025.

48. *Eppel S.* LAS&T: Large Shape And Texture Dataset. — 05.2025. — DOI: 10.5281/zenodo.15453634. — URL: https://doi.org/10.5281/zenodo.15453634.

49. Qwen2. 5-vl technical report / S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang [и др.] // arXiv preprint arXiv:2502.13923. — 2025.

50. Qwen3 technical report / A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv [и др.] // arXiv preprint arXiv:2505.09388. — 2025.

51. RADIOv2. 5: Improved baselines for agglomerative vision foundation models / G. Heinrich, M. Ranzinger, Y. Hongxu, Y. Lu, J. Kautz, A. Tao, B. Catanzaro, P. Molchanov // Proc. CVPR. Т. 2. — 2025. — С. 6.

52. SegmATRon: Embodied adaptive semantic segmentation for indoor environment / T. Zemskova, M. Kichik, D. Yudin, A. Staroverov, A. Panov // Neurocomputing. — 2025. — Т. 638. — С. 130169.

53. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features / M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa [и др.] // arXiv preprint arXiv:2502.14786. — 2025.

54. Tulip: Towards unified language-image pretraining / Z. Tang, L. Lian, S. Eisape, X. Wang, R. Herzig, A. Yala, A. Suhr, T. Darrell, D. M. Chan // arXiv preprint arXiv:2503.15485. — 2025.

55. How to Segment Wires, Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot / L. Vetoshkin, O. Matykina, M. Kichik, U. Izmesteva, D. Yudin // Hoses, and Other Obstructive Objects in the Traversable Area of a Mobile Robot. —.