Распознавание физических свойств объектов трехмерных сцен по их RGB-изображениям

Кичик М.Г.
kichik.mg@phystech.edu

Магистратура Центра когнитивного моделирования МФТИ, 2 курс

2025 – Научно-исследовательский семинар

Мотивация работы и постановка задачи

Распознавание физических свойств объектов – полезная в практическом смысле задача, которая редко встречается в научных публикациях. Существующие работы посвящены, как правило, созданию новых датасетов, причем доля искусственных данных, в частности созданных с помощью симуляторов, существенно больше доли настоящих изображений. На данном этапе главная задача – определение материала объекта, потому что на основе этой характеристики естественным образом можно оценить физические свойства.

Цель работы заключается в исследовании и усовершенствовании существующих подходов для определения физических свойств объектов. Рассматриваются как классические подходы в компьютерном зрении (модели семантической сегментации), так и визуальные языковые модели (VLM). Применение первой группы подходов естественно для поставленной задачи. Применение VLM кажется уместным, поскольку языковые модели используют информацию не только из визуального контекста, но и из "собственных" знаний о мире: это может улучшить точность определения материалов.

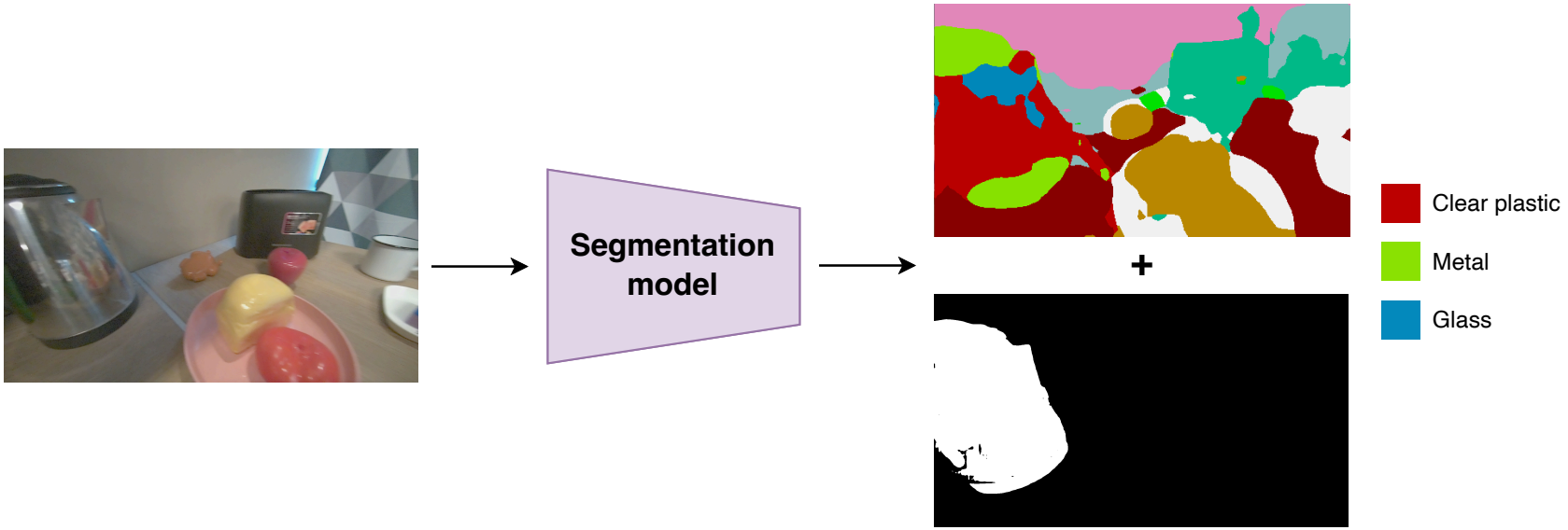
Датасеты и симуляторы с данными по материалам

Датасет	Год публикации	Детали
MINC [2]	2015	Датасет содержит натуральные фотографии, размеченные бинарными масками и кликами. Использовано 23 категории материалов.
3D Scene Graph [1]	2019	В датасете представлены 3D сцены из Gibson Environment, у каждого объекта в описании есть поле, соответствующее набору материалов, из которых сделан объект. Для описания материалов использовано 23 категории (те же, что и в датасете MINC).
3DSSG [10]	2020	Датасет содержит семантические графы сцен, материалы объектов кодируются 15 категориями. Использование датасета зависит от установки 3RScan [Wald2019RIO].
Hypersim [8]	2021	Датасет содержит 77000 изображений из 416 внутренних помещений, сгенерированных в симуляторе. Остутствует фиксированный набор категорий по материалам объектов, но в мета-информации есть названия материалов.
3D-FUTURE [4]	2021	В датасете представлено 20000 3D сцен внутренних помещений, представлено 15 категорий для описания материалов.
DMS [9]	2022	Датасет состоит из 44000 натуральных фотографий, размеченных семантическими масками. Разметка содержит 52 категории материалов.
MINERVAS [7]	2022	Симулятор, предоставляющий в числе прочих функций возможность конфигурировать материалы объектов сцены.
SGFRONT	2023	Датасет содержит семантические графы сцен, информация о материалах представлена в форме сопоставления: объект сделан из того же материала, что и другой объект, или нет.

Датасет DMS, примеры

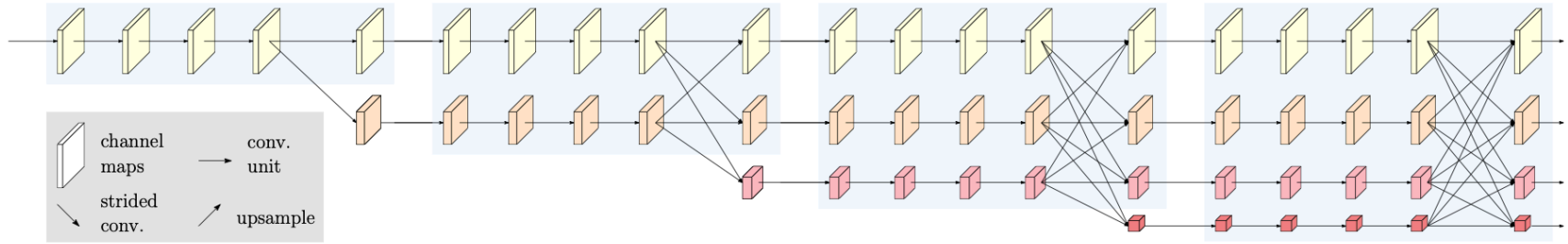


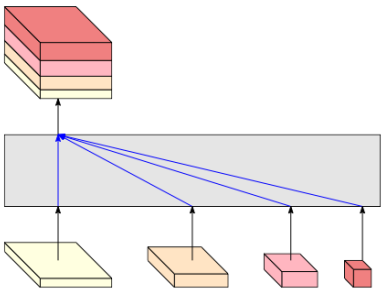
Схема работы сегментационных моделей



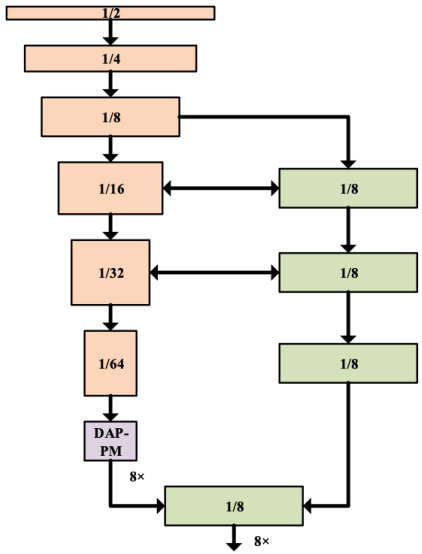
Сегментаторы: DDRNet [6], HRNet [12], Mask2Former [3]

HRNet

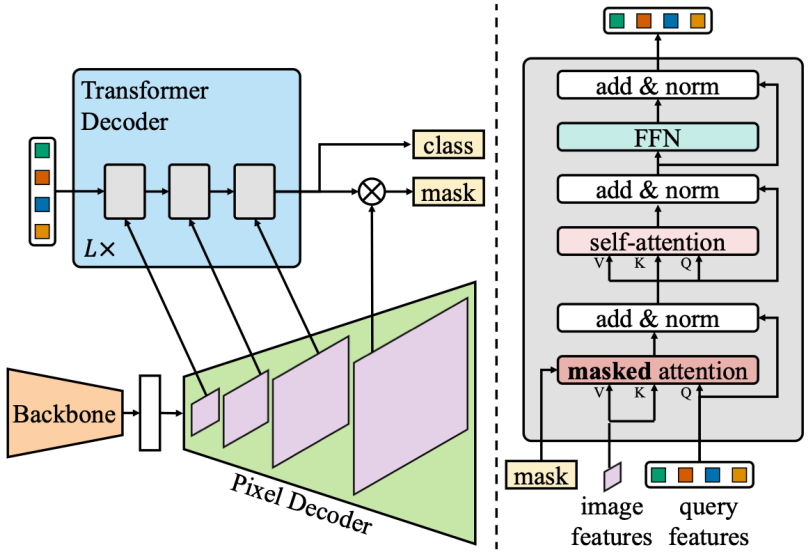




DDRNet



Mask2Former



Результаты сегментационных моделей, обученных на DMS

Метрики подсчитаны на DMS-val.

Модель	mIoU	mAcc	Время инференса, мс	Потребляемая память, МБ
DMS	**43.23**	**59.67**	19	1694
HRNet	36.55	46.71	80	2056
DDRNet	33.35	43.36	26	1878
Mask2Former	35.69	43.25	76	2804

Качественные результаты сегментационных моделей

Для сравнения работы моделей были использованы фотографии кухни.

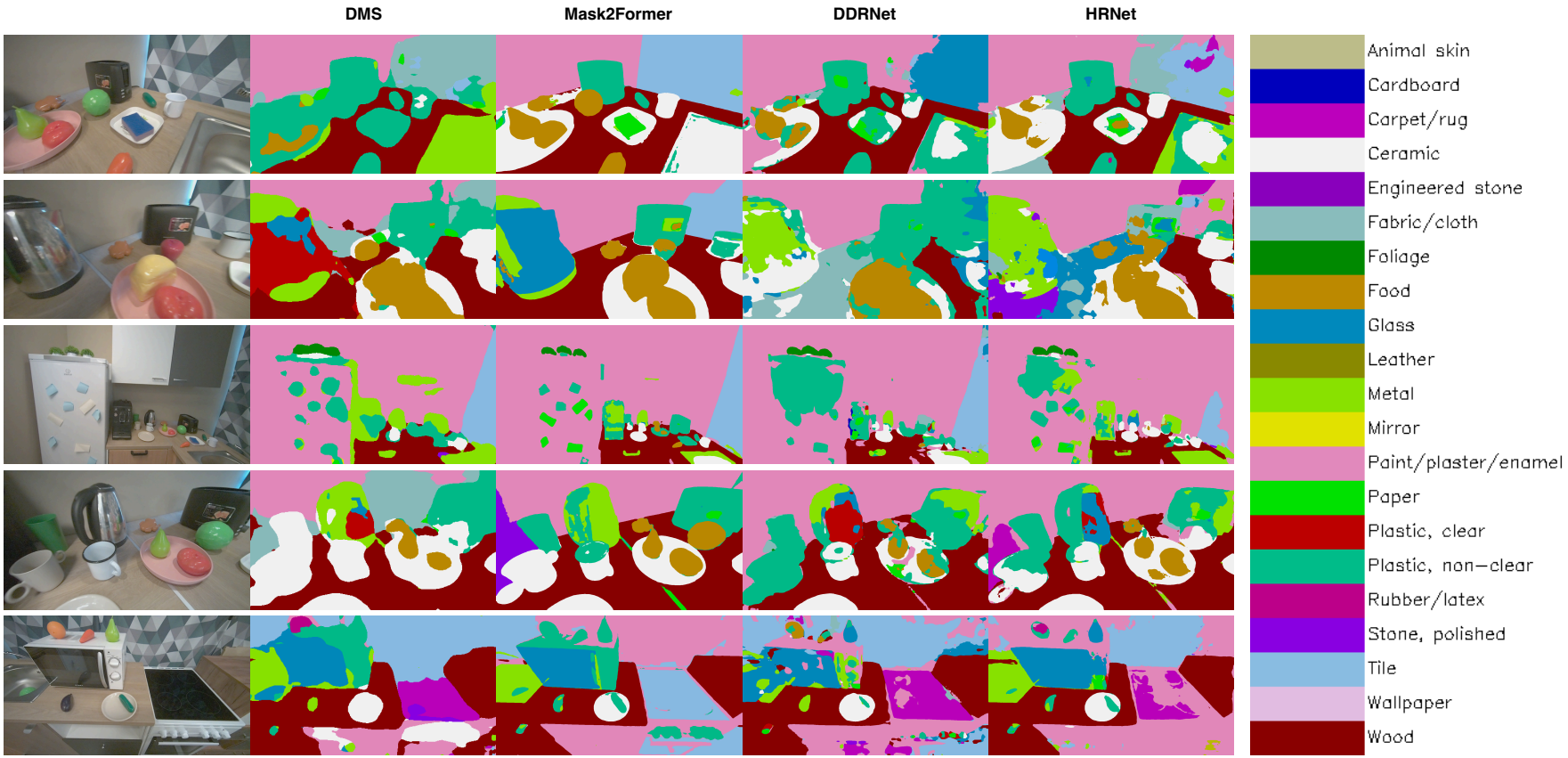


Схема работы VLM



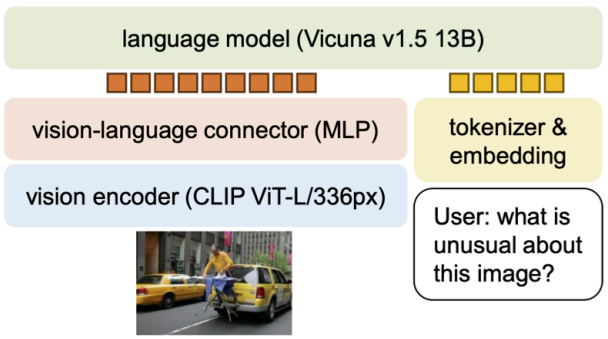


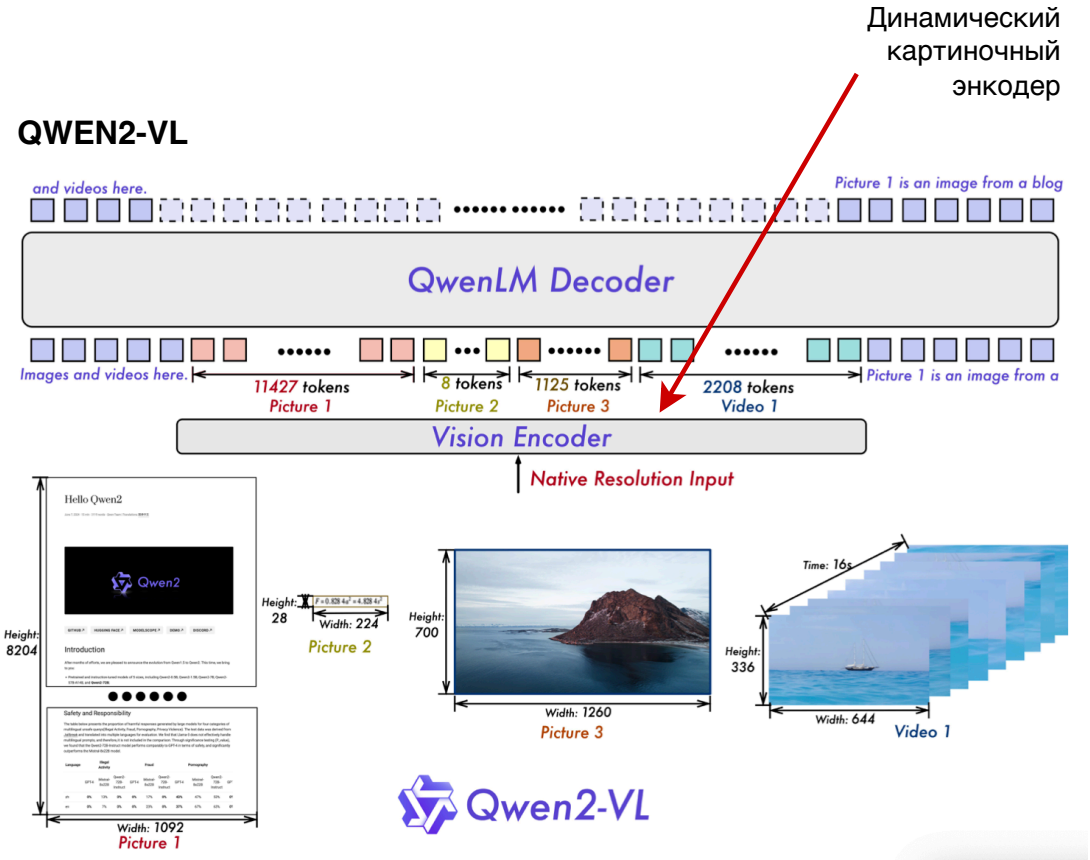
You receive a bounding box [0.0, 0.11, 0.32, 0.86],
which represents [xmin, ymin, xmax, ymax].
Fit you description up to 4–5 words.
Name the material of the object in the bounding box.
Answer in the form of [object_name], [object_material].

Visual Language model

Tea kettle, metal

VLMs: LLaVa 1.5 [5], QWEN2-VL [11]





Оценка качества работы VLM

Для оценки VLM применялась метрика, схожая с Precision:

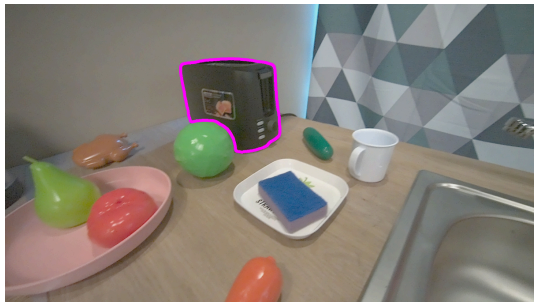
$$\text{Precision} = \frac{\Sigma_{\text{all objects}} \text{predicted material} \in \text{caption}}{N_{\text{all objects}}}$$

Результаты валидации VLM

Валидация происходила на 5 размеченных изображениях кухни (55 объектов).

Модель	Детали эксперимента	Точность	Время инференса, с	Потребляемая память, МБ
DMS	-	0.536	0.02	1694
HRNet	-	0.500	0.08	2056
QWEN2-VL-2B	Обычные изображения, размер 1440x810.	0.607	3.22	11756
QWEN2-VL-2B	Обычные изображения, размер 1080x607.	0.607	1.62	7274
QWEN2-VL-2B	Обычные изображения, размер 640x360.	0.589	0.55	5800
QWEN2-VL-2B	Обрезанные изображения, добавлены границы в 25% от ширины изображения, передаются координаты ограничивающего прямоугольника.	0.411	0.97	17176
QWEN2-VL-2B	Обрезанные изображения, добавлены границы в 25% от ширины изображения, не передаются координаты ограничивающего прямоугольника.	0.482	1.04	17176
QWEN2-VL-7B	Обычные изображения, размер 1440x810.	0.625	5.26	23232
QWEN2-VL-7B	Обычные изображения, размер 1080x607.	0.607	2.80	18898
QWEN2-VL-7B	Обычные изображения, размер 640x360.	0.536	1.20	17410
LLaVa1.5-7B	Обычные изображения, размер 1440x810.	0.518	0.51	14456
LLaVa1.5-7B	Обрезанные изображения, добавлены границы в 25% от ширины изображения, передаются координаты ограничивающего прямоугольника.	0.428	0.47	14584
LLaVa1.5-7B	Обрезанные изображения, добавлены границы в 25% от ширины изображения, не передаются координаты ограничивающего прямоугольника.	0.554	0.26	14384

Качественные результаты VLM (LLaVa1.5-7B и QWEN2-VL-2B)



LlaVa: Toaster, Plastic
QWEN2-VL: Plastic



LlaVa: Refrigerator, stainless steel
QWEN2-VL: White plastic



LlaVa: Sink, stainless steel
QWEN2-VL: Metal



LlaVa: Green apple, plastic
QWEN2-VL: Plastic



LlaVa: A yellow cake on a pink plate.
QWEN2-VL: Plastic



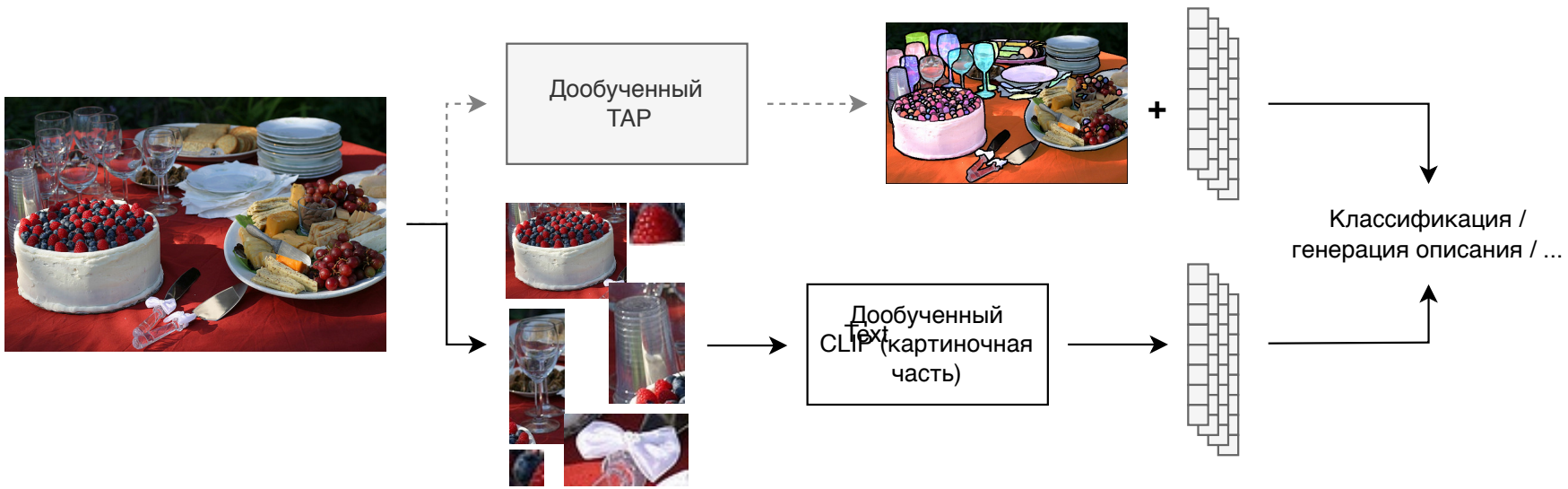
LlaVa: Tea kettle, metal
QWEN2-VL: Glass

Промежуточные выводы (2024)

- Результаты моделей семантической сегментации на DMS приблизительно соответствуют одному и тому же значению по метрикам mIoU и mACC (около 35 и 44 пунктов). Такое качество кажется недостаточно высоким для использования моделей в реальных задачах и при этом скорее связано с особенностями датасета, а не самих моделей.
- VLM показывают достаточно хорошие результаты на маленьком "кухонном" датасете, но у текущего подхода существует ряд недостатков. VLM достаточно медленные, в особенности по сравнению с моделями сегментации, и иногда вместо корректного ответа они галлюцинируют: это влияет не только на корректность ответа и возможность его обработать, но и на время, которое зависит от длины сгенерированного текста. Недостатком подхода является сложность оценки неопределенности ответа модели, потому что работа происходит с уже сгенерированным текстом.

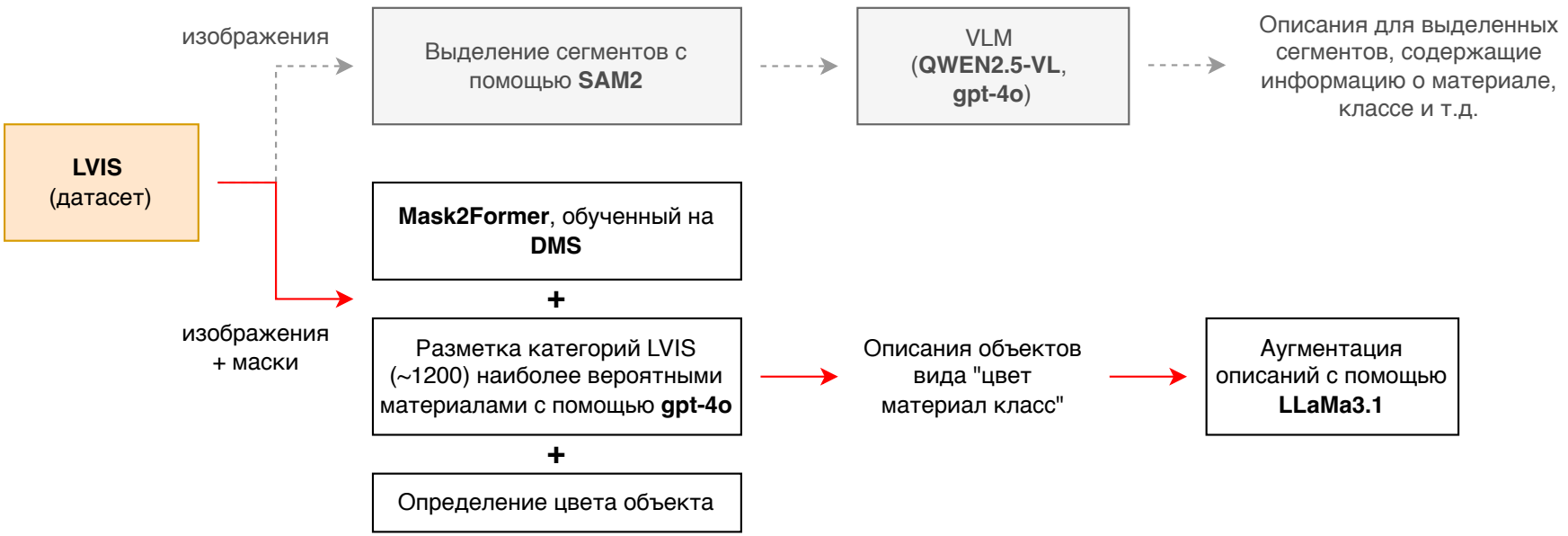
Распознавание материалов объектов на изображении с открытым словарем

Задача – разработать метод распознавания материалов, в котором словарь (набор классов) не будет фиксированным. Алгоритм должен определять материалы для заданных объектов и быть масштабируемым на новые домены / наборы классов без полного переобучения.



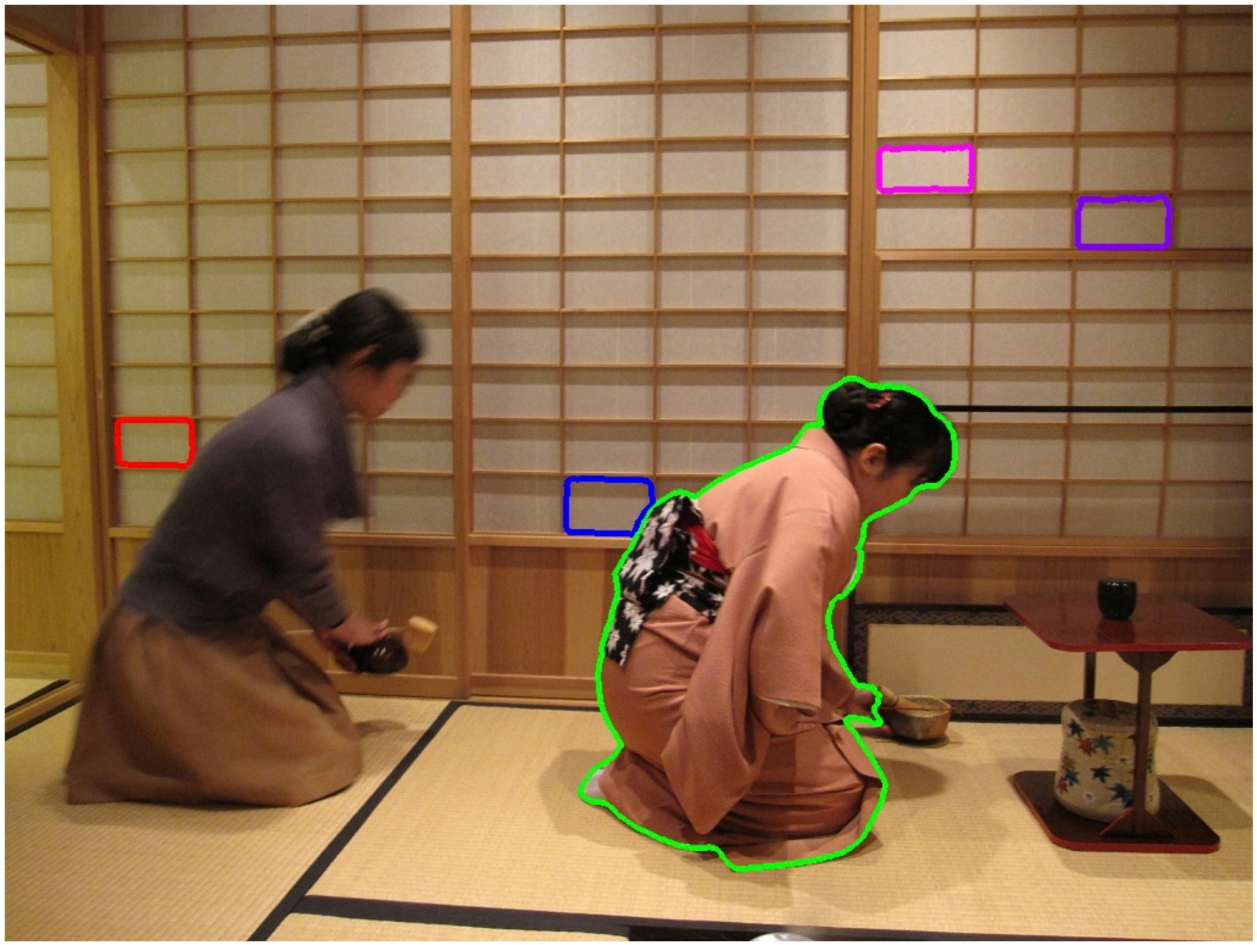
Пайплайн сборки датасета

Датасет собирали на основе LVIS. В тренировочной выборке 1M масок, в тестовой – 50K.
Для создания датасета рассматривали комбинацию SAM2 + gpt4-o, но отказались от такого подхода из-за сомнительного качества и скорости работы (около 10с на 5 масок). Сейчас пайплайн включает использование Mas2Former, обученного на DMS, и QWEN2.5-VL.



Разметка с помощью VLM

Для разметки выбирали по 5 масок, обводили контур на изображении и соответствующим образом формулировали промпт для модели.



Разметка gpt-4o-mini

structured	*free form*
A small red wooden box. It is hard, smooth, and strong.	The box has an elegant design, often used for storage or as a decorative piece. It adds a touch of charm to its surroundings.
A medium blue rectangular box. It is hard, smooth, and strong.	This box features a simple yet stylish appearance and is typically used for organizing items or gifts, making it practical in various settings.
A small brown wooden table. It is hard, smooth, and stable.	The table serves as a functional surface, perfect for placing items like tea or decorations. Its compact size makes it versatile for different spaces.
A medium-sized pink ceramic bowl. It is hard, smooth, and strong.	The bowl has a delicate and graceful design, often utilized for serving or displaying food, adding a vibrant touch to table settings.
A small violet tea cup. It is hard, smooth, and sturdy.	The tea cup is elegantly crafted, designed for enjoying hot beverages, and often features intricate patterns that enhance its aesthetic appeal.

Описания в основном не соответствуют действительности и сильно зависят от цвета обводки.

Разметка gpt-4o

structured	*free form*
A small, brown, wooden wall panel. It is hard, smooth, strong.	This is part of a wooden panel that forms part of the wall in a traditional Japanese room.
A small, beige, paper or fabric shoji screen panel. It is soft, smooth, fragile.	This is a part of the shoji screen, a traditional Japanese sliding door made of wood and paper or fabric.
A medium, pink, fabric kimono. It is soft, smooth, strong.	The person is wearing a traditional pink kimono, which is typical attire in Japanese tea ceremonies.
A small, beige, paper or fabric shoji screen panel. It is soft, smooth, fragile.	Another section of the shoji screen, which is a fundamental feature in traditional Japanese interior design.
A small, beige, paper or fabric shoji screen panel. It is soft, smooth, fragile.	This is yet another part of the shoji screen, contributing to the room's serene and elegant atmosphere.

Описания хорошие, но модель ошибается на мелких сегментах или если на изображении очень много объектов (например, фото магазина одежды).

Разметка QWEN2.5-VL (3B)

structured	*free form*
A small, rectangular box with a glossy finish. It is hard, smooth, and sturdy.	A shiny, compact storage container that fits neatly into the corner.
A medium-sized, rectangular box with a matte finish. It is hard, smooth, and sturdy.	A solid, non-glossy storage box that holds items securely.
An outline of a person wearing traditional attire. The outline is not a physical object but rather a visual representation of the person's shape.	A green line tracing the silhouette of a person dressed in traditional clothing.
A small, rectangular box with a glossy finish. It is hard, smooth, and sturdy.	A shiny, compact storage container that fits neatly into the corner.
A small, rectangular box with a glossy finish. It is hard, smooth, and sturdy.	A shiny, compact storage container that fits neatly into the corner.

Модель не всегда поддерживает заданную структуру ответа и часто ошибается.

Примеры из получившегося датасета (DMS + аугментация с LLaMa)





A deep grey metal handle with a sturdy construction.



A grey telephone made from plastic materials.



A deep grey wooden desk with a rich finish.

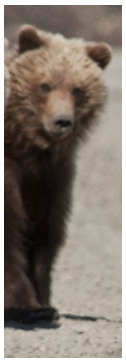


Warm-toned yellow glass wine glass.

A muted grey lampshade made of plastic material.

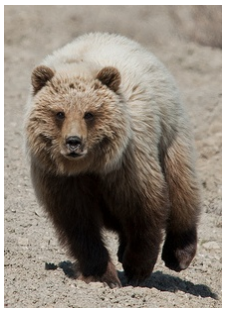






A deep olive-coloured bear with a thick, sturdy fabric-like fur.

A muted grey bear with soft fabric fur.



Обучение CLIP: метрики

$$\textbf{Recall@K} = \frac{1}{N}\sum_{i=1}^{N} 1(\text{rank}_i \leq K)$$

$$\textbf{Mean Cosine Similarity} = \frac{1}{N}\sum_{i=1}^{N} \cos(\theta_i), \text{ где } \cos(\theta_i) = \frac{\mathbf{x}_i \cdot \mathbf{y}_i}{\|\mathbf{x}_i\|\|\mathbf{y}_i\|}$$

$$\textbf{MRR (Mean Reciprocal Rank)} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\text{rank}_i}$$

Выбор гиперпараметров для обучения CLIP

CLIP обучали на обрезанных по ограничивающему прямоугольнику изображениях. Выбор гиперпараметров проводили при обучении на тестовой выборке. Исследовали влияние таких факторов, как аугментация описаний, размытие фона и обучение без мелких объектов (10 пкс по меньшему измерению). Для основного эксперимента выбрали конфигурацию *аугментированные описания + размытие фона + без мелких объектов*.

модель	тренировка	аугментированные кэпшены	блюр	без мелких объектов	recall@1	recall@5	recall@10	mean cosine similarity	MRR
ViT-B/32					0,316	0,585	0,704	0,242	0,447
ViT-B/32		да			0,308	0,571	0,687	0,235	0,435
ViT-B/32			да		0,299	0,569	0,688	0,242	0,429
ViT-B/32		да	да		0,288	0,547	0,669	0,233	0,416
ViT-B/32				да	0,376	0,669	0,777	0,248	0,512
ViT-B/32		да		да	0,372	0,658	0,762	0,241	0,504
ViT-B/32			да	да	0,362	0,643	0,763	0,247	0,496
ViT-B/32		да	да	да	0,354	0,632	0,746	0,24	0,485
ViT-B/32	да				0,303	0,685	0,84	0,825	0,472
ViT-B/32	да	да			0,295	0,69	0,847	0,817	0,47
ViT-B/32	да		да		0,288	0,682	0,851	0,858	0,464
ViT-B/32	да	да	да		0,274	0,678	0,848	0,849	0,452
ViT-B/32	да			да	0,282	0,679	0,852	0,839	0,453
ViT-B/32	да	да		да	0,271	0,678	0,843	0,834	0,448
ViT-B/32	да		да	да	0,298	0,71	0,879	0,863	0,479
ViT-B/32	да	да	да	да	0,299	0,708	0,877	0,861	0,477
ViT-B/32	да	random-micro	да	да	0,287	0,693	0,866	0,86	0,465
ViT-B/32	да	random-macro	да	да	0,299	0,721	0,877	0,858	0,482

Результаты обучения CLIP

	Recall@1	Recall@5	Recall@10	MCS	MRR
CLIP [ViT-B/32], default	0.346	0.629	0.745	0.239	0.480
CLIP [ViT-B/32], after train [mcs]	0.358	0.795	0.926	**0.913**	0.543
CLIP [ViT-B/32], after train [mrr]	**0.403**	**0.820**	**0.927**	0.888	**0.580**

Дальнейшие планы

- Обучить клип вместе с классификационной головой для определения материалов (по фиксированному набору концептов);
- Обучить CLIP с динамическим картиночным энкодером (во избежание апскейлинга маленьких изображений до 224x224);
- Обучить CLIP с замороженным текстовым энкодером (например T5);
- Собрать небольшой качественно размеченный с помощью SAM2 + gpt4-o для дообучения моделей;
- Обучить (дообучить) TAP на существующем датасете;

Спасибо за внимание!

kichik.mg@phystech.edu

Библиография I

[1] Iro Armeni и др. «3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera». В: *Proceedings of the IEEE International Conference on Computer Vision*. 2019.

[2] Sean Bell и др. «Material Recognition in the Wild with the Materials in Context Database». В: *Computer Vision and Pattern Recognition (CVPR)* (2015).

[3] Bowen Cheng, Alexander G. Schwing и Alexander Kirillov. «Per-Pixel Classification is Not All You Need for Semantic Segmentation». В: 2021.

[4] Huan Fu и др. «3d-future: 3d furniture shape with texture». В: *International Journal of Computer Vision* (2021), с. 1—25.

[5] Haotian Liu и др. *Improved Baselines with Visual Instruction Tuning*. 2024. arXiv: 2310.03744 [cs.CV]. URL: https://arxiv.org/abs/2310.03744.

[6] Huihui Pan и др. «Deep Dual-Resolution Networks for Real-Time and Accurate Semantic Segmentation of Traffic Scenes». В: *IEEE Transactions on Intelligent Transportation Systems* (2022).

Библиография II

[7] Haocheng Ren и др. *MINERVAS: Massive INterior EnviRonments VirtuAl Synthesis*. 2022. arXiv: 2107.06149 [cs.CV]. URL: https://arxiv.org/abs/2107.06149.

[8] Mike Roberts и др. «Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding». В: *International Conference on Computer Vision (ICCV) 2021*. 2021.

[9] Paul Upchurch и Ransen Niu. «A Dense Material Segmentation Dataset for Indoor and Outdoor Scene Parsing». В: *European Conference on Computer Vision (ECCV)*. 2022.

[10] Johanna Wald и др. «Learning 3D Semantic Scene Graphs from 3D Indoor Reconstructions». В: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020.

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

[12] Yuhui Yuan, Xilin Chen и Jingdong Wang. «Object-Contextual Representations for Semantic Segmentation». В: (2020).