

Курс «Мультимодальные БЯМ»

Тема №3
«3D LLM»

Антон Конушин



Понятие о 3D зрении

Что такое компьютерное зрение?



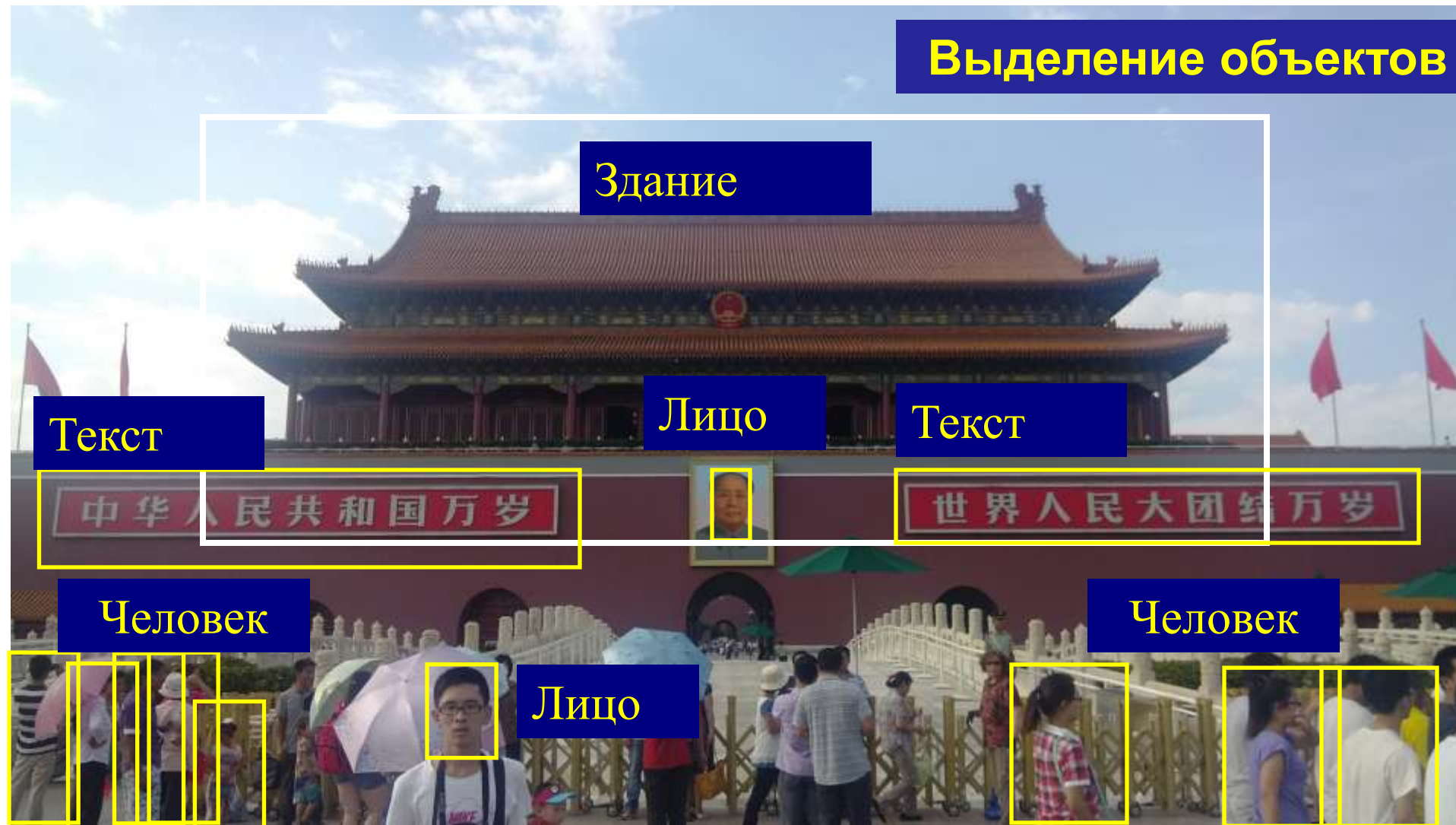
Задача зрения: понять, что находится на изображении

Компьютерное зрение: построение компьютерной модели системы зрения

Компьютерное зрение – часть области искусственного интеллекта (AI)

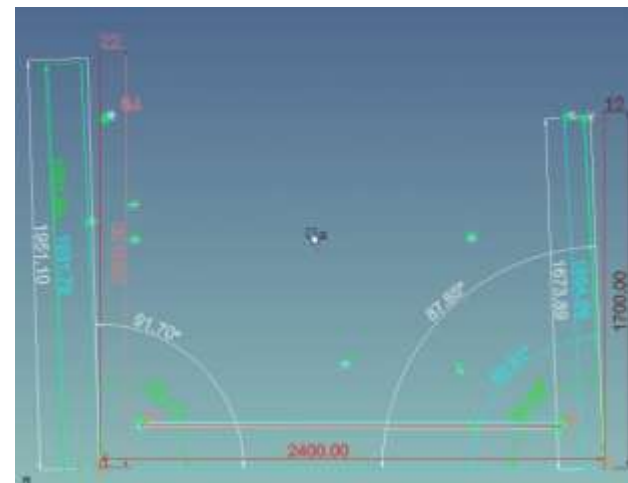
Тест Тьюринга для компьютерного зрения: Ответить на любой вопрос про изображении, на который может ответить человек.

Что и где? (Семантическое зрение)



Необходимо определить, есть ли на изображении объекты заданного типа и если да, то определить их положение

Какой формы? (Метрическое зрение)



Фотограмметрия

Source: Pollefeys et al.

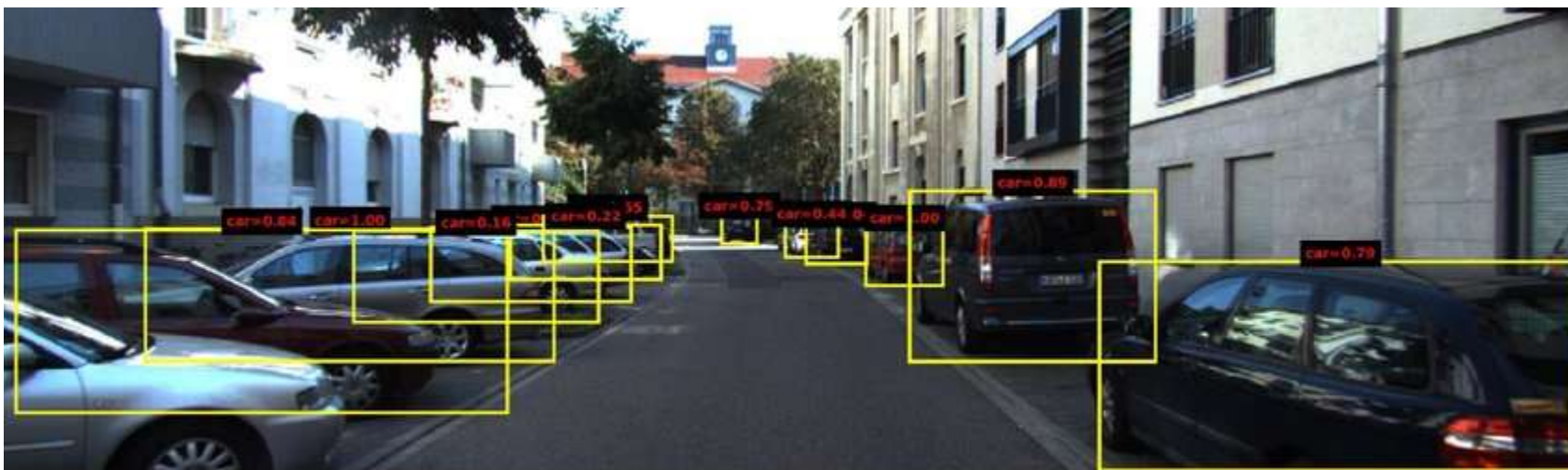


3-х мерная реконструкция по изображениям

3D зрение



3D зрение = Метрическое зрение + Семантическое зрение



2D детекция



3D детекция



Как получать информацию о 3D?

3D реконструкция по изображениям



Набор изображений



Трёхмерная (3D) модель

Пример задачи: построение 3D модели реального объекта по набору фотографий из интернета

Примеры реконструкций



Dubrovnik, Croatia.

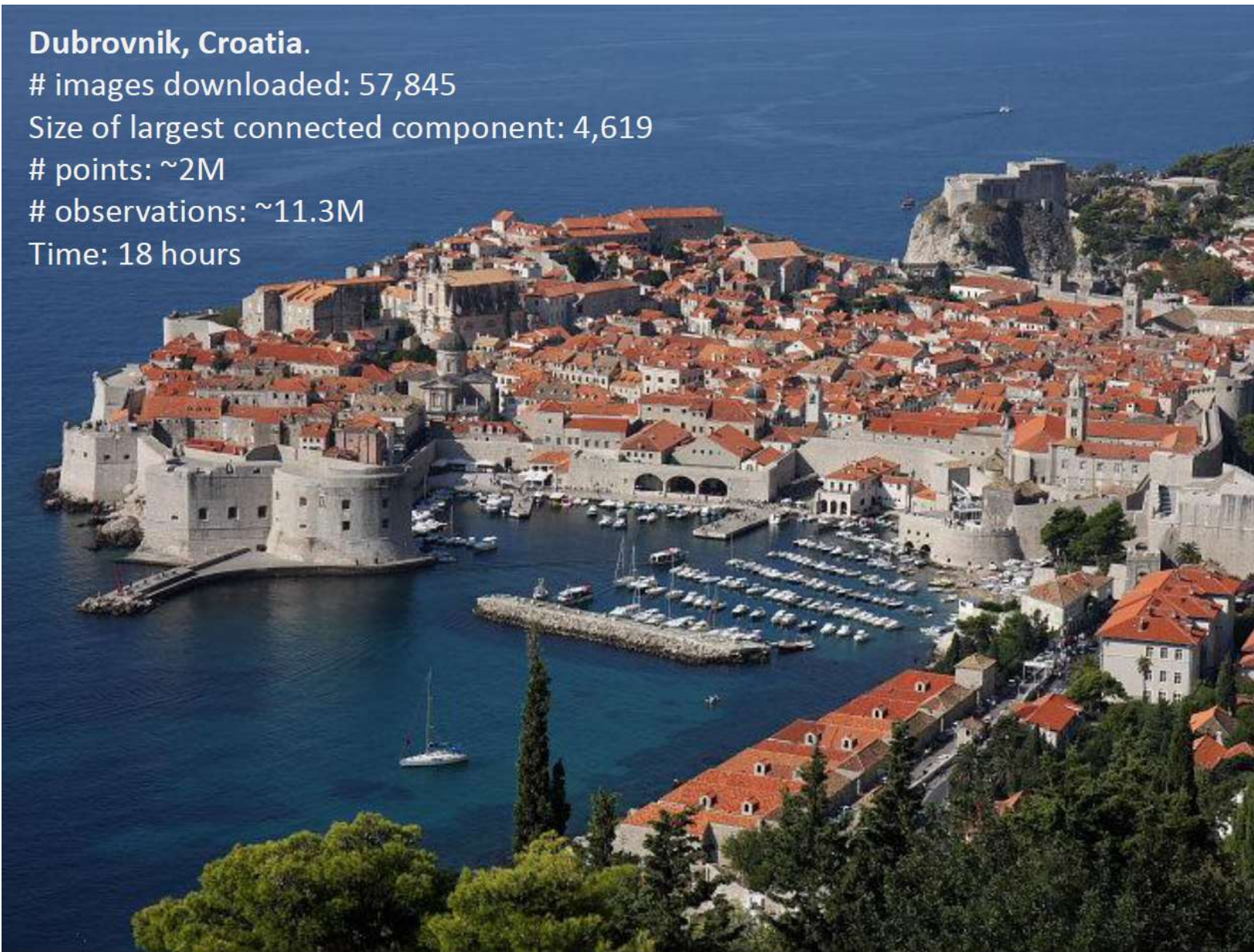
images downloaded: 57,845

Size of largest connected component: 4,619

points: ~2M

observations: ~11.3M

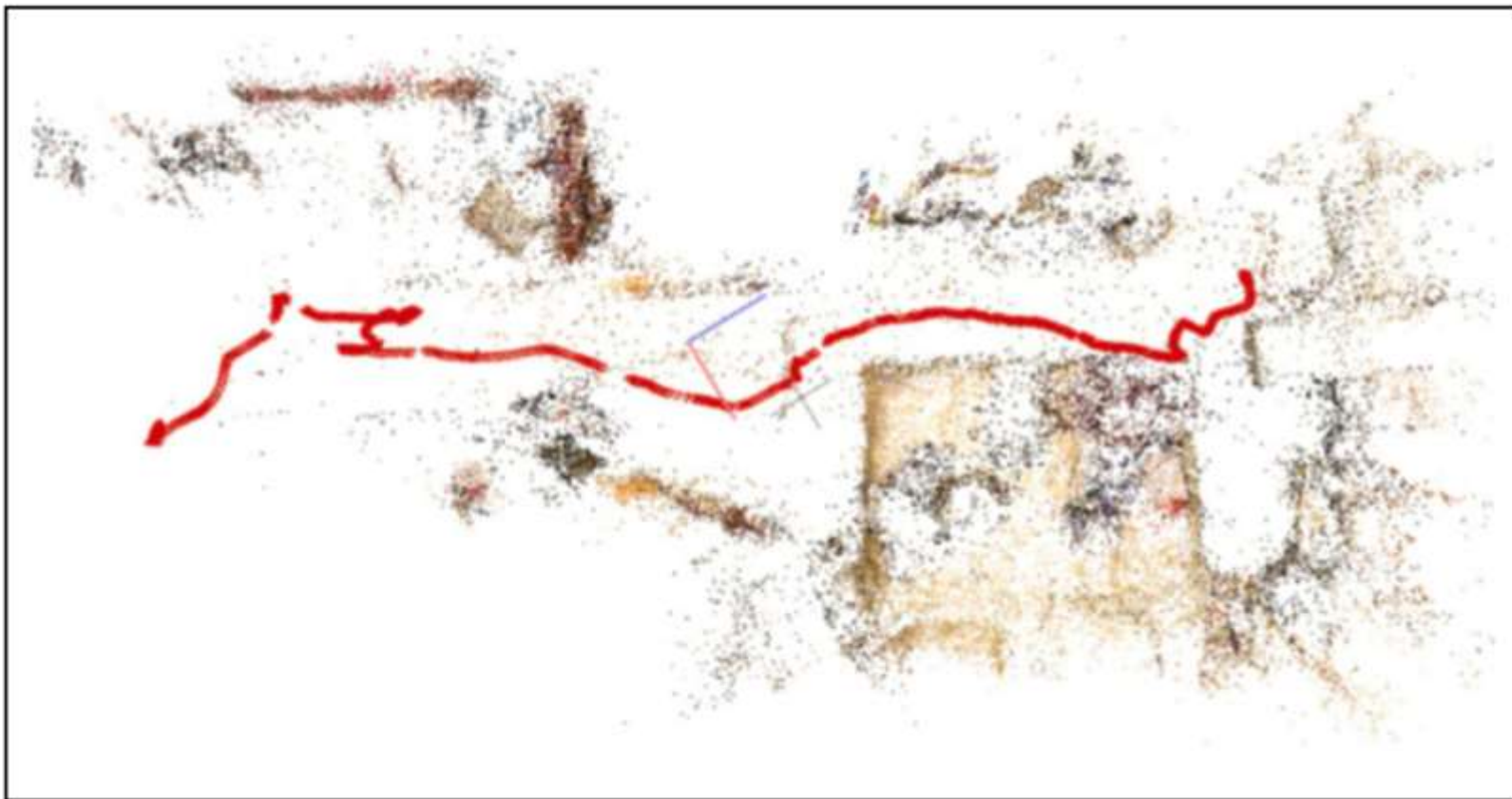
Time: 18 hours



Дубровник



Источники данных – RGB изображения



Пример реконструкции с помощью методов Structure-from-Motion (SfM)

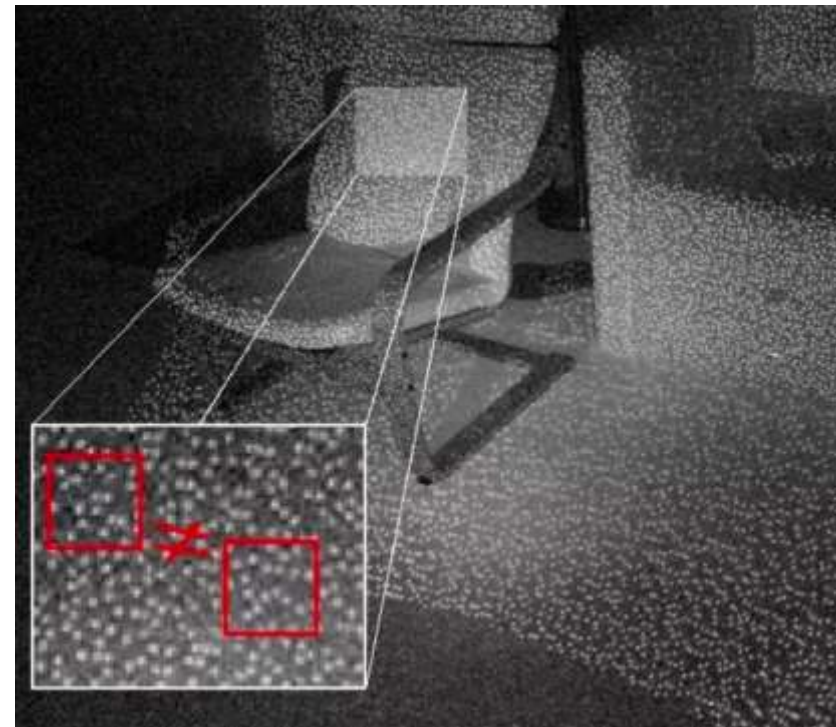
Сенсоры глубины (Depth Camera)



Технология компании [PrimeSense](#), лицензированная Microsoft и реализованная в камере [Kinect](#) for Xbox 360 ("Project Natal")



Kinect for Xbox One &
Kinect for Windows - Time
of Flight камеры



"Умная" структурная подсветка в виде набора пятен по хитрому шаблону. Форма пятна также анализируется для оценки глубины и нормалей.

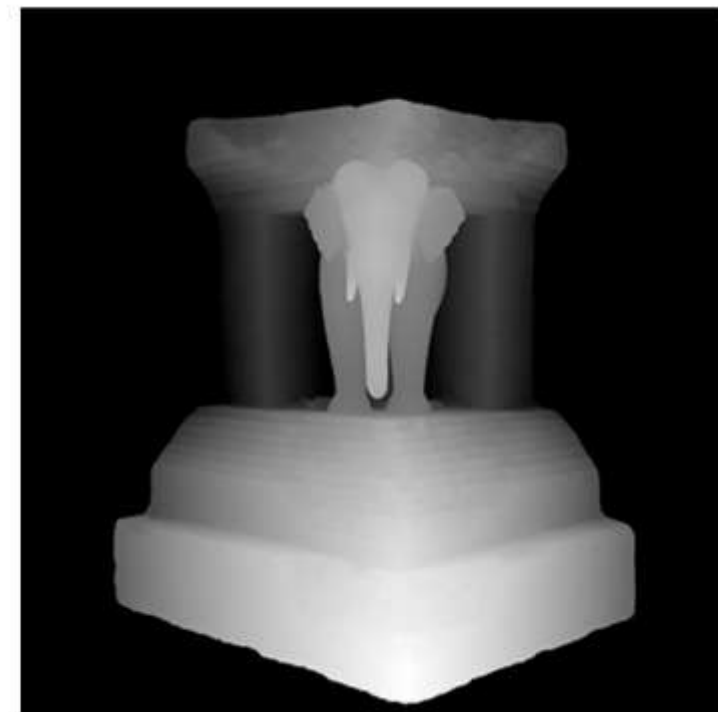
Карты глубины (Depth Maps)



Сенсоры глубины



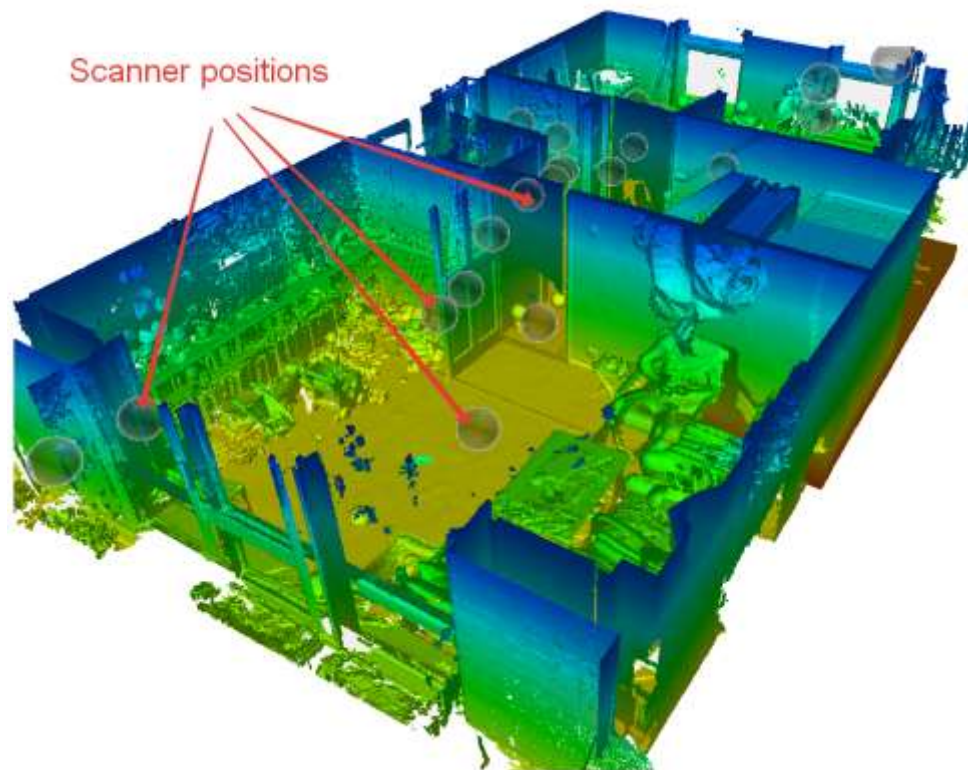
Изображение



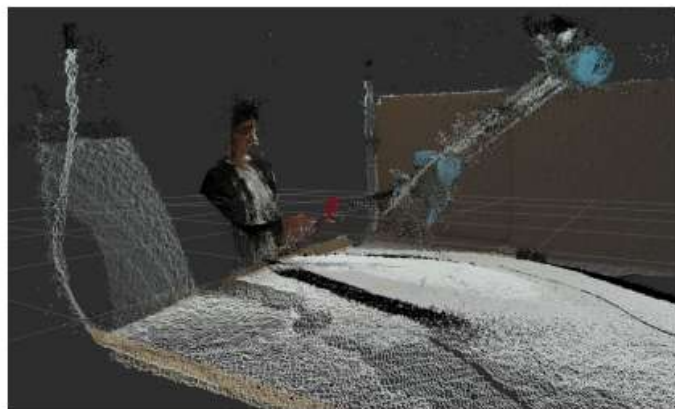
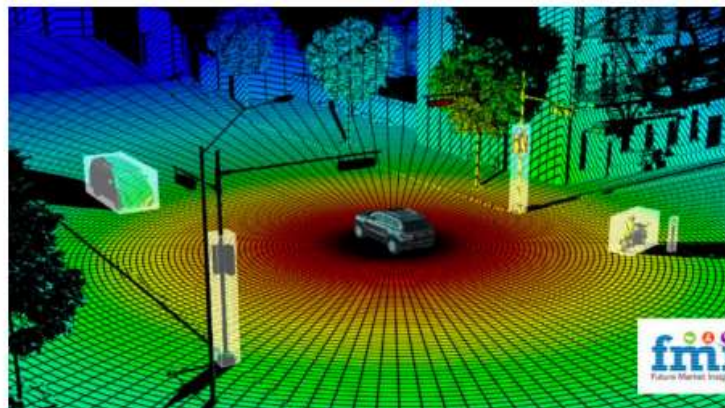
Карта глубины

- Карта глубины – изображение с 1м каналом, в котором записаны расстояния до ближайшего объекта сцены
- Часто объединяют с RGB изображением, получается RGBD изображение
- Можем использовать обычные методы обработки изображений

Лазерные сканеры



Облака точек





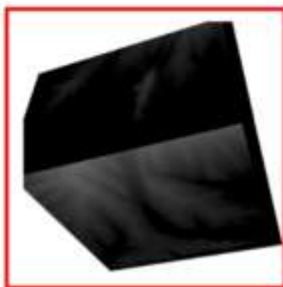



- Облака точек обычно получаются (“acquired”) процессом, подобным получению изображений
- Например, lidar строит «панораму» окружающей сцены
- Но бывает удобно «забыть» про исходную природу 3д точек
- Каждому пикселю изображения с глубиной мы можем сопоставить 3Д точку. Множество пикселей даст облако точек
- Мы можем объединить информацию с разных ракурсов в одно общее облако точек



3D задачи и датасеты

3D Captioning



						
BLIP2 (1-view)	a green baseball ball	a green monster with big eyes	a black and white image of a cube	a black and white image of a room with furniture	a fish swimming in the water	a sculpture of a monkey's head
GPT4-Vision (1-view)	a baseball	a stylized figure looks like a cat ...	a folded piece of fabric , possibly a black garment	... a couch , coffee table, and possibly a television	a bird in mid-flight	a representation of an orangutan's face
Cap3D (8-views)	... a mix of a frog , teddy bear , and monster with big eyes and red or pink eyes .		a small room featuring a bed, desk with computer, chair, bookshelf, and bathroom sink .		3D sculpture of an orangutan head with a hat , accompanied by a pelican and a fish .	
GPT4-Vision (28-views)	...an amorphous creature with one eye , a partially visible set of sharp teeth...		A minimalist-style interior room featuring a sofa , low table, shelving with books, and a standalone cabinet.		...with dynamic, fluid-like distortions altering ... appears to be made of a smooth material .	
Ours (6-views)	A turquoise-colored 3D character with stylized feline features, prominent eyes, whimsical pink facial markings, and zipper-like teeth detail.		Monochromatic 3D model of a room interior featuring a two-door cabinet, a central coffee table, a bed with pillows, a side chair, a trash bin, and a wall-mounted bookshelf with books, all rendered in grayscale with solid and shaded surfaces suggesting lighting effects.		A 3D sculpture of a gorilla's head with realistic texturing in shades of gray and brown, featuring detailed facial attributes, expressive eyes, and a distinctly modeled facial structure.	

Классификация объектов – слишком тривиально, поэтому начнём с 3D captioning!

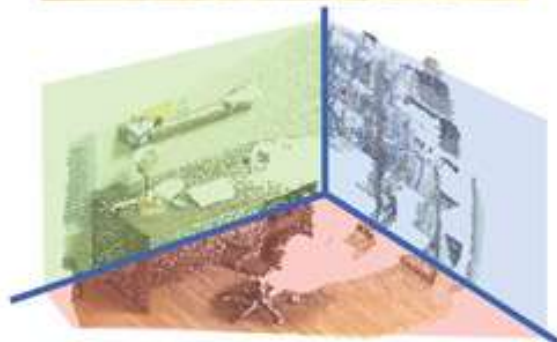
SUN RGB-D Benchmark (2015)



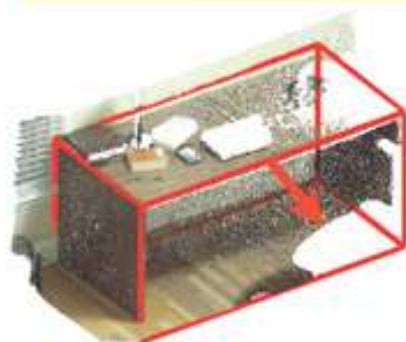
Scene Classification



Semantic Segmentation



Room Layout



Detection and Pose

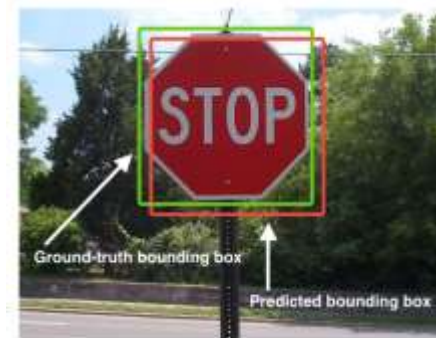
Total Scene Understanding

4 разных камеры, 10,000 RGB-D изображений (размер как у PASCAL VOC), 58,657 3D bounding boxes

3D Detection



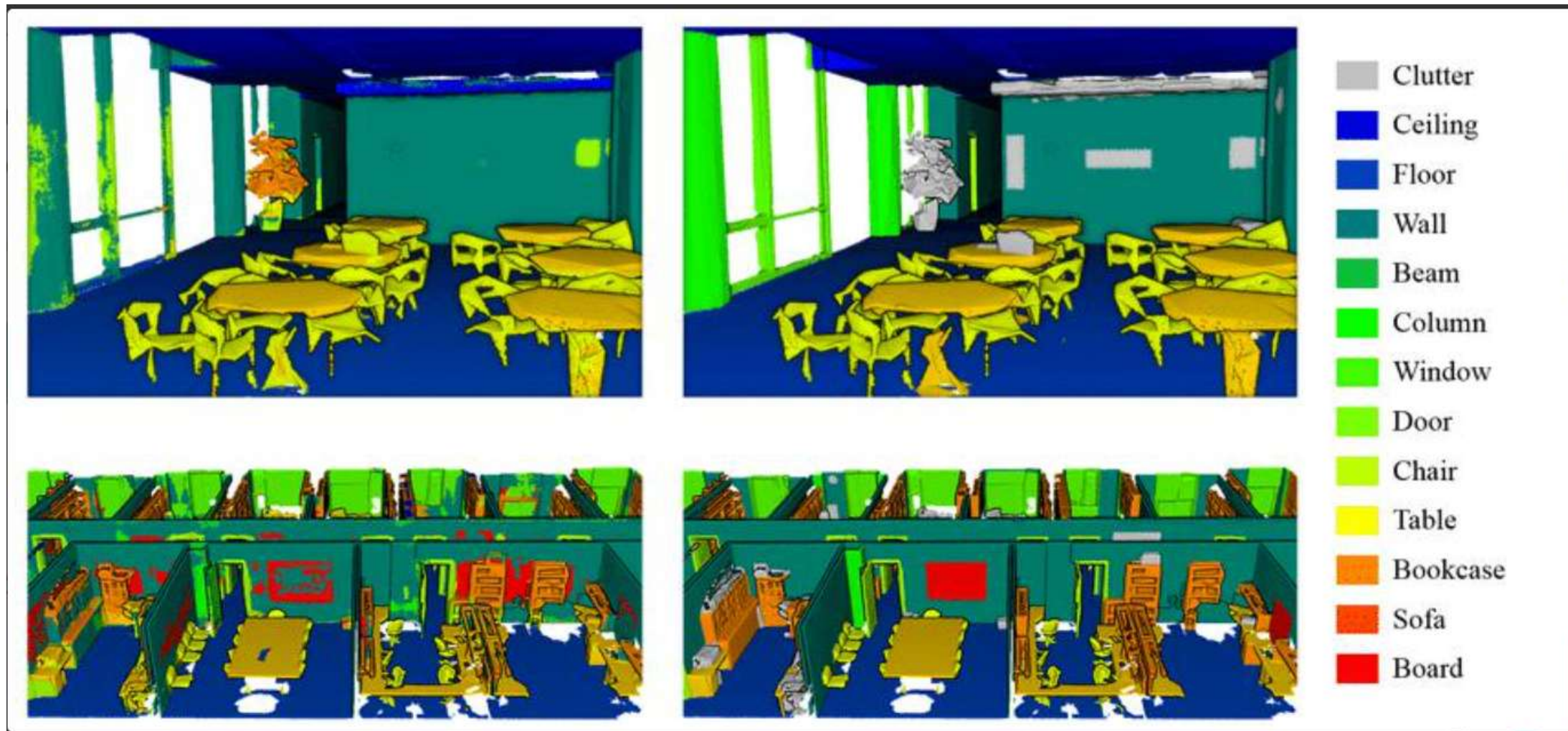
Total Scene Understanding



$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$
A diagram illustrating the calculation of Intersection over Union (IoU). It shows two overlapping blue squares. The intersection of the two squares is highlighted in a darker blue. Below the squares, the formula for IoU is shown, with the intersection area corresponding to the 'Area of Overlap' and the combined area of both squares corresponding to the 'Area of Union'.

- Результат - 3D bounding box
- Вариант 1 - Axis Aligned bounding box
- Вариант 2 - Oriented bounding box (повернутый в плоскости земли)
- Оценка совпадения с GT - Intersection over Union (IoU)
- Оценка качества детекции - mAP @ 0.25-0.5 IoU

3D Семантическая сегментация

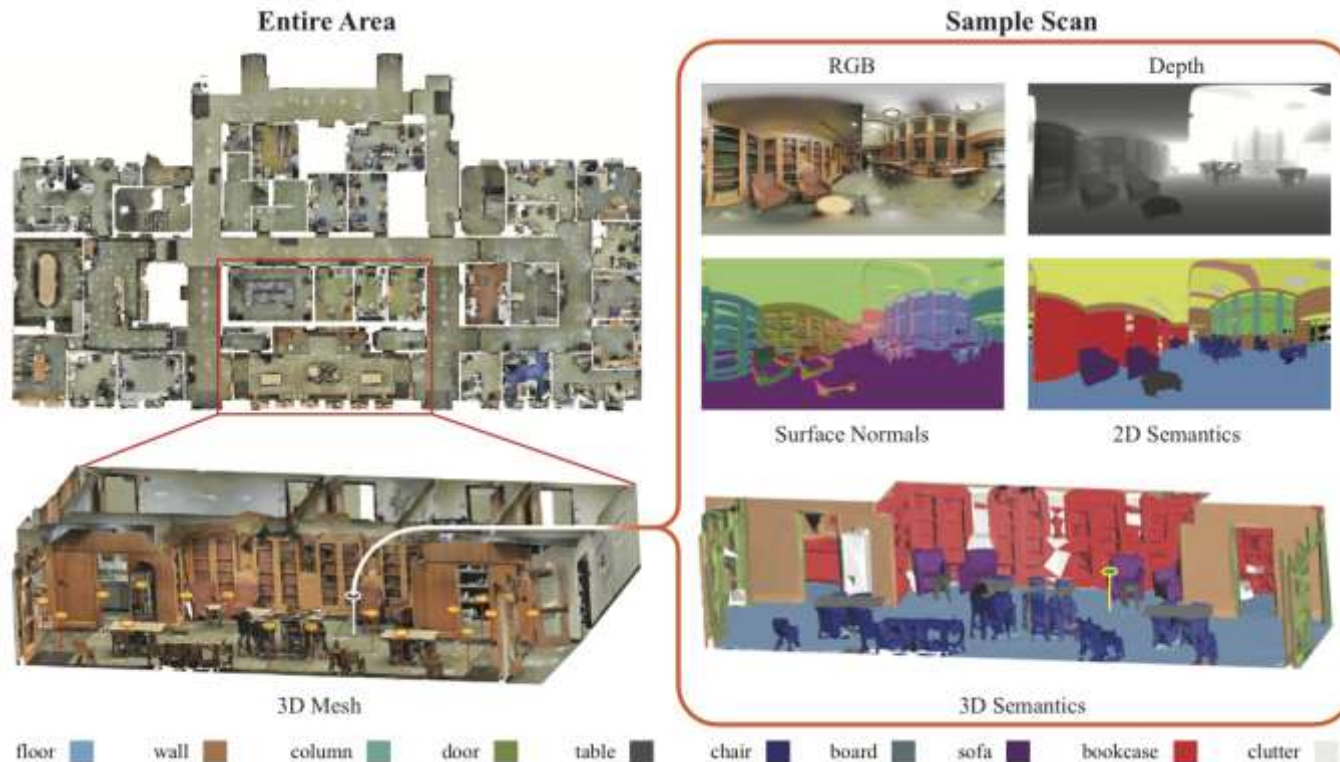


3D Instance & Panoptic Segmentation

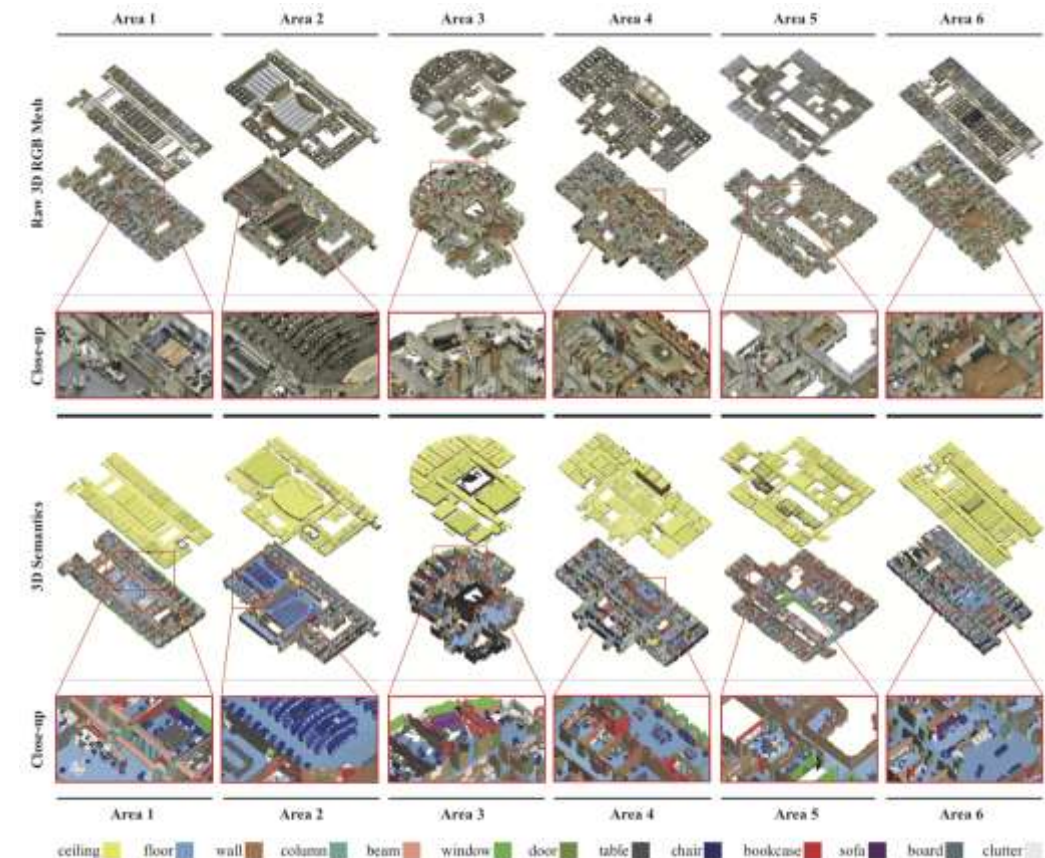


- Instance segmentation -
разметка облака
точек метками отдельных
экземпляров
объектов (instance)
- Panoptic segmentation -
двойная разметка облака
точек и метками экземпляров,
и метками классов объектов

Stanford 3D Indoor Scene Dataset (S3DIS, 2016)

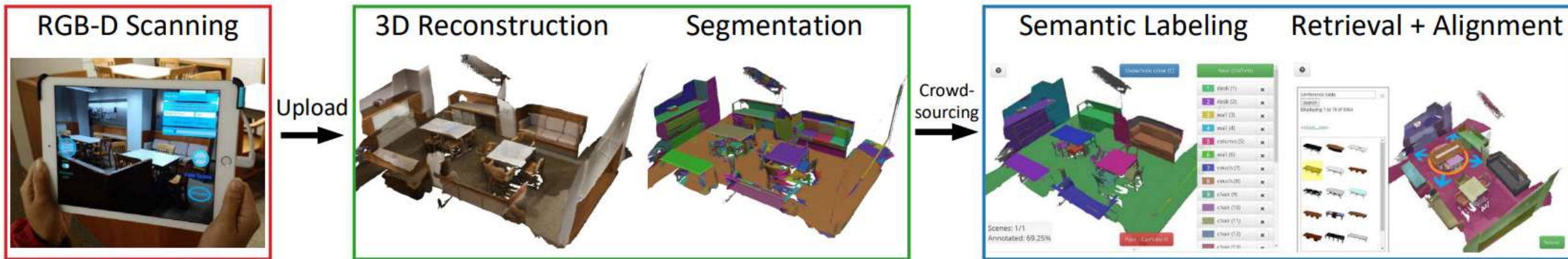


3D реконструкция сцен, облака точек, разметка семантики



6 больших сцен

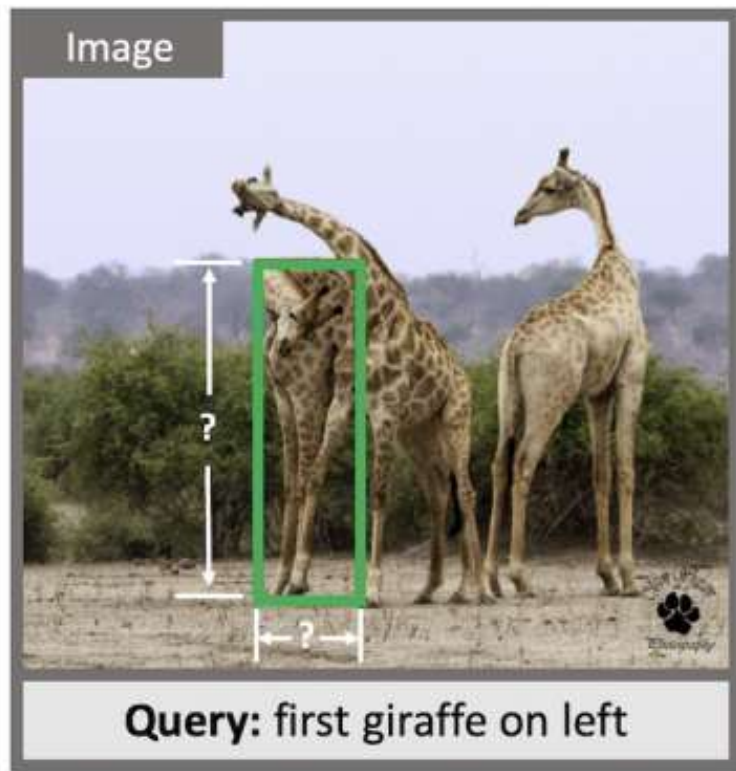
ScanNet (2017)



- 1513 сканов, 700+ помещений, более 2.5M RGBD кадров
- Все виды данных выложены, point cloud, семантическая и instance разметка, и т.д.
- Используется и как бенчмарк, и как обучающий набор для большинства методов

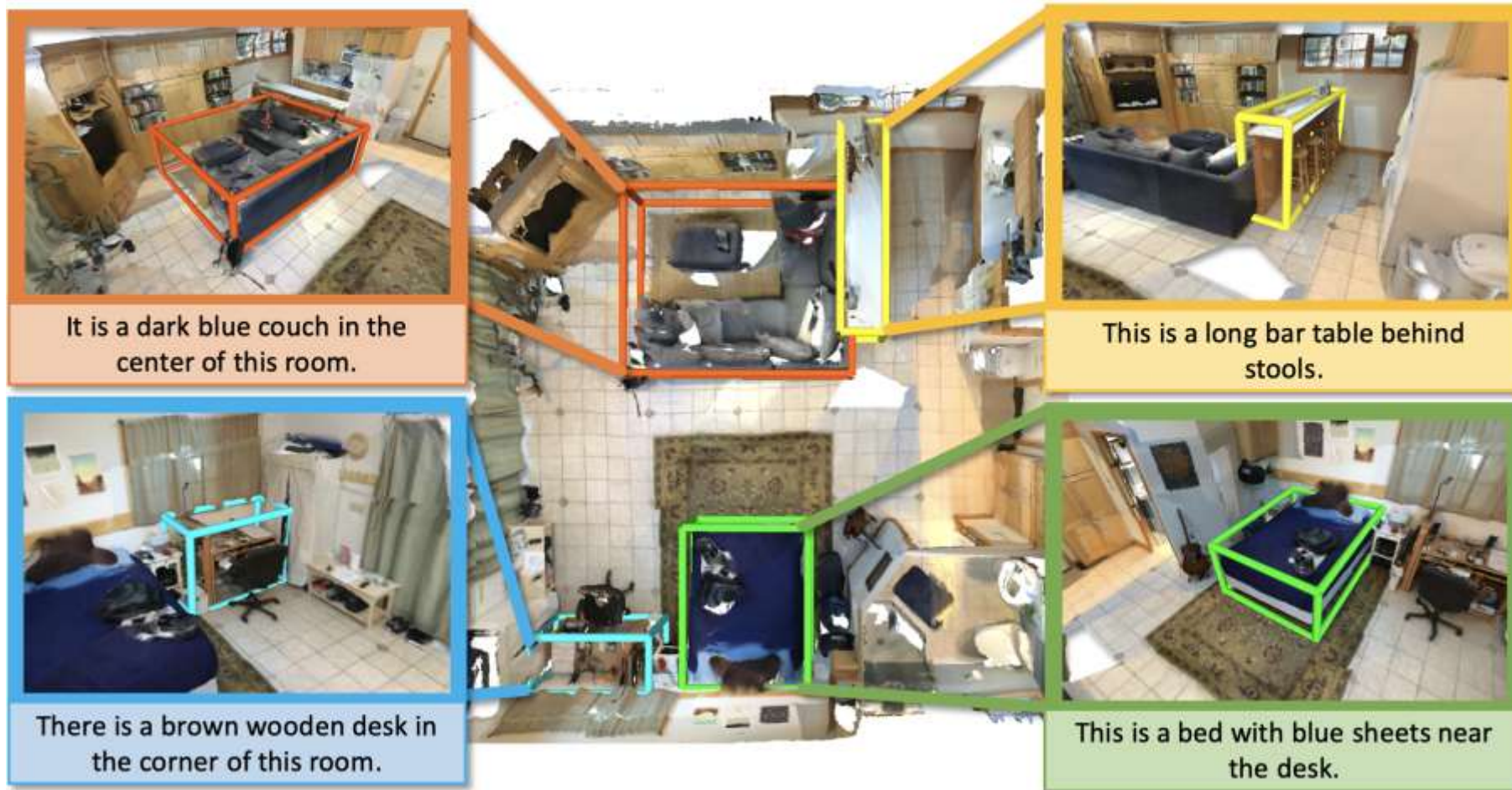
<http://www.scan-net.org/>

Задачи 2D и 3D Visual Grounding



Локализация объекта по текстовому описанию
(мультимодальная задача)

ScanRefer (2020)



- 51,583 текстовых описаний для 11,046 объектов
- 800 ScanNet сцен
- Разметка вручную через Amazon Turk
- Проверка студентами – нужно найти объект по описанию

3D Question Answering



Question + 3D-Scan



Q. Where is the medium sized blue suitcase laid?

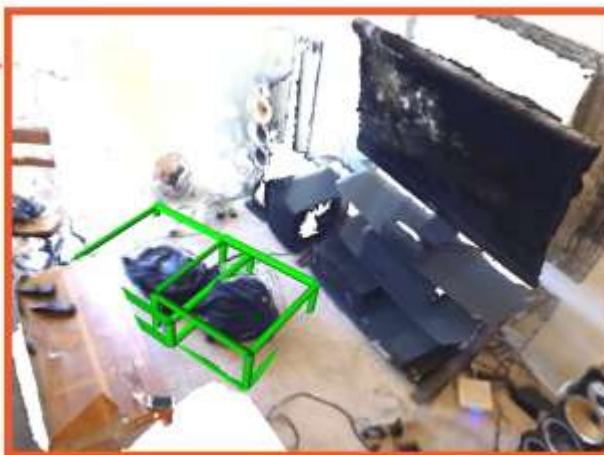
Answer + 3D-Bounding Box



A. in front of right bed



Q. What is sitting on the floor between the tv and the wooden chair?



A. 2 black backpacks

- Расширение 3D Visual Grounding
- Нужно отвечать на вопросы о свойствах объектов, относительном расположении, находить группы объектов и т.д.
- Ответ бывает текстовый, и текст + 3D bbox объектов

ScanQA



Underspecified questions

- Q: What is in the corner?
- Several objects at corners!
- Q: What color is the chair?
- Three chairs at the scene!

Valid questions

- Q: What is over the chair beneath the blackboard?
- Answer: jacket
- Q: What color is the office chair next to the desk with a monitor?
- Answer: green

3D-QA Datasets	Type	Question Collection	Answer Collection	Environment	Photorealistic	# 3D Scenes
IQUAD	Interactive	Template-based	Template-based	AI2THOR	No	30 rooms
EQA	Navigation	Template-based	Template-based	House3D	No	588 scenes
MP3D-EQA	Navigation	Template-based	Template-based	Matterport 3D	Yes	144 floors
MT-EQA	Navigation	Template-based	Template-based	House3D	No	588 scenes
ScanQA dataset	3D Scan	AutoGen+HumanEdit	Human	ScanNet	Yes	800 rooms

Table 1. Comparison of 3D question-answering datasets.

Split	# Question	# Unique Question	# 3D Scenes
Train	25,563	20,546	562
Val	4,675	4,306	71
Test w/ objects	4,976	4,552	70
Test w/o objects	6,149	5,484	97
Total	41,363	32,337	800

Table 2. ScanQA dataset statistics.

ScanNet + ScanRefer + доразметка

SceneVerse (2024)



- Объединение датасетов + генерация различных текстовых описаний по шаблонам и LLM
- Можно обучить мощную модель для 3D Visual Grounding

<https://scene-verse.github.io/> (ECCV 2024)

Суммируем задачи

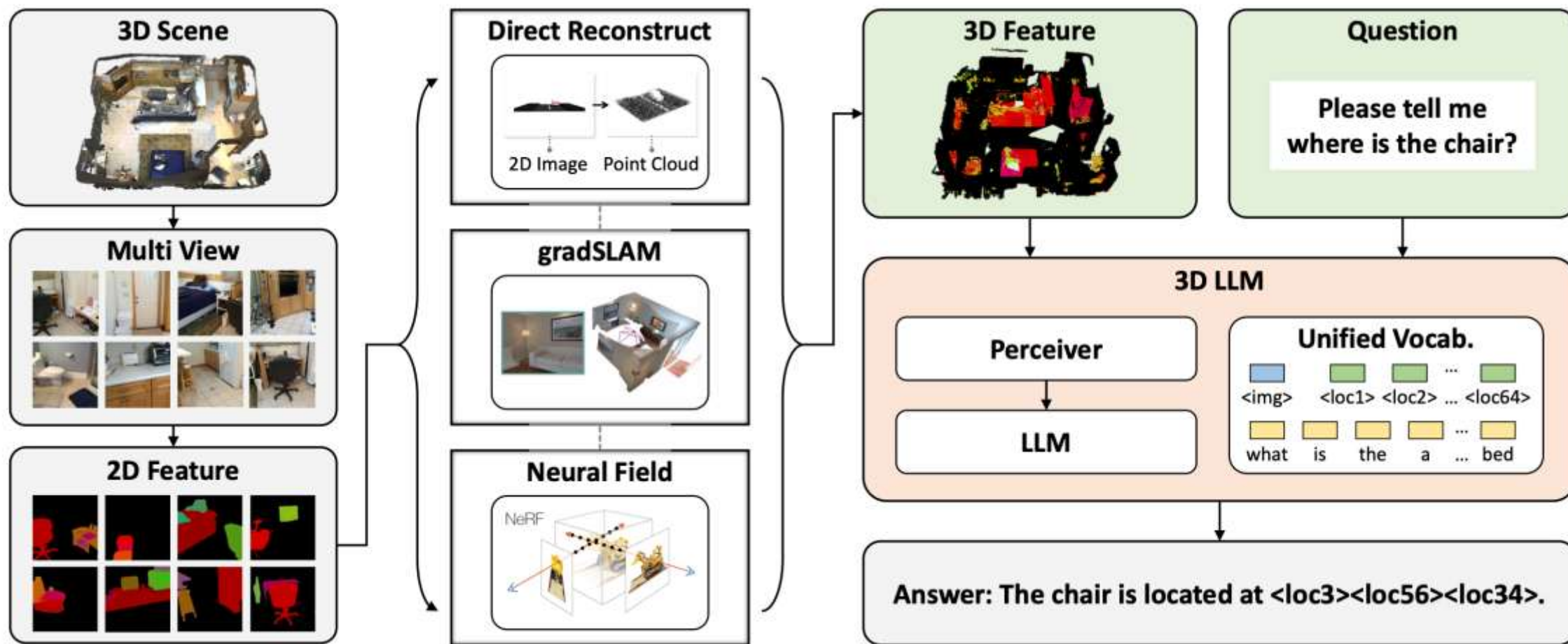


- 3D classification & 3D object / scene captioning
- 3D detection
- 3D semantic segmentation, 3D instance segmentation, 3D panoptic segmentation
- 3D visual grounding как развитие задачи 3D detection
- 3D Question Answering



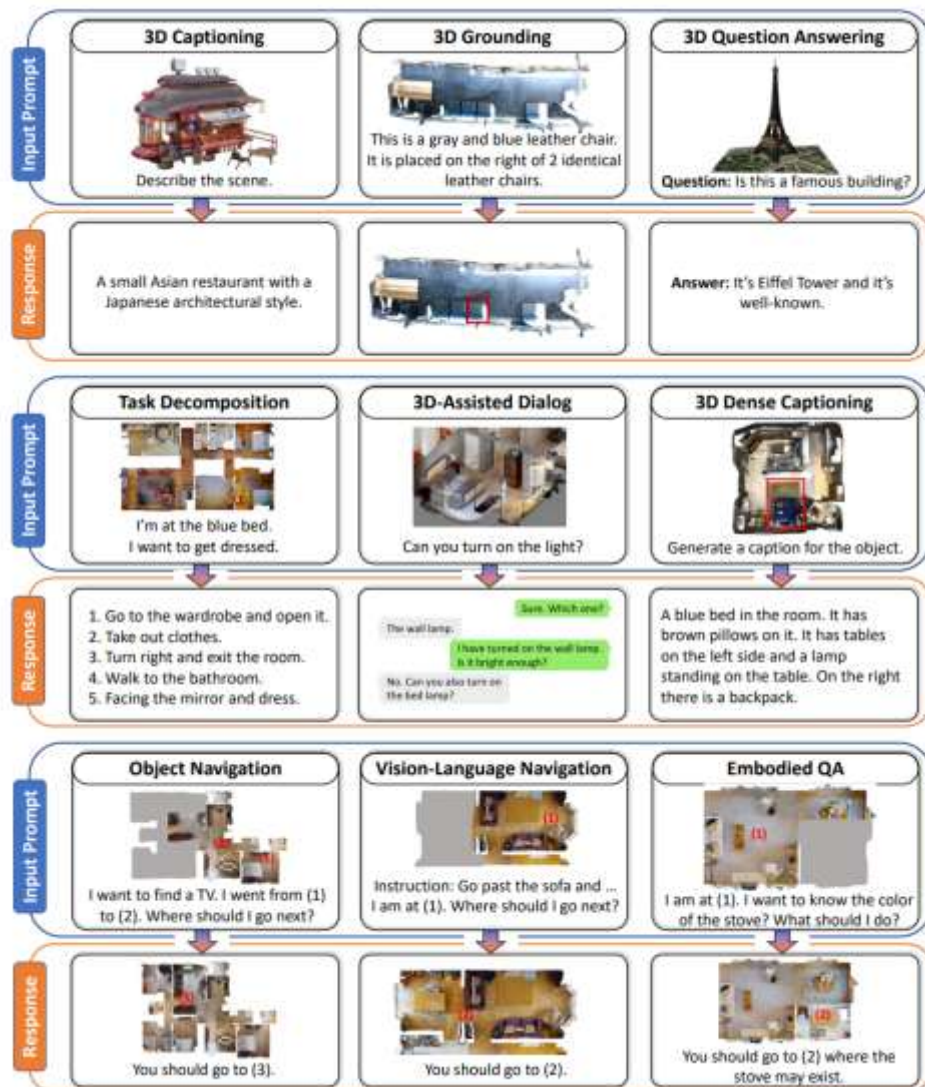
Визуальные большие языковые модели

3D-LLM (Первая LLM для 3D)



Важно – проецирование CLIP признаков в 3D, дополнение фич positional embeddings & location tokens

Генерация данных для модели



Box-Demonstration-Instruction based Prompting

Bounding-Box Context:

Room 1: table: [0.4, 0.1, 0.1], chair: [0.4, 0.3, 0.4]... Room2: desk: [0.5, 0.2, 0.3]

Instruction (Shortened Version):

You are an **AI visual assistant** / **conversation generator** / **question generator** in the room. You need to **generate a high-level task and its low-level actions** / **4-10 round conversations between a human and a robot assistant** / **question-answer pairs** about the scene.

Demonstration: For Room 1:

<Task Decomposition> High-Level Task: make up. Low-Level Actions: 1. go to the cabinet 2. take out cosmetics from the cabinet 3. bring the cosmetics to the table with mirror 4. open the lamp 5. make up

<3D-Assisted Dialog> Human: Bring me a toothbrush. Robot: Sure! Where is it?. Human: Maybe in the sink. Robot: I have arrived at the sink and find no toothbrush. I found an closed bathroom cabinet. Is it possible to contain what you need?...

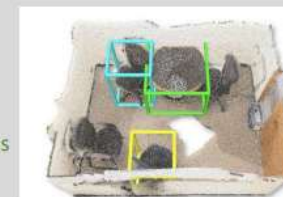
<QA> Question: Is there a table between the chair and the tv? Answer: Yes. Generate similar responses for Room 2.

Response : For Room 2:

<Task Decomposition> High-Level Task: Get ready for work. Low-Level Actions: 1. go to the toilet 2. Wash hands in the sink cabine. 3. Gog to the mirror 4. Brush teeth and wash face 5. Turn on the lamp 6. Leave the room through the door

<3D-Assisted Dialog> Human: Turn on the light. Robot: Sure! Which light?. Human: The wall lamp. Robot: I have turned it on. Is it bright enough? Human: No. Are there other lights? Robot: I see a desk lamp.

<QA> Question: Viewing the tv from the desk, is there a guitar on the right? Answer: Yes.



ChatCaptioner based Prompting



ChatGPT
Captioner

What can you see from the far right?

Can you tell me what's on the desk?

Anythng else close to the desk?

What can you see from the near left?

What's the color of the bed?

...



BLIP
Answerer

Chair, desk

A laptop

A curtain

A bed

Blue

Images from different views



Response: Here's the description of the scene: This room has a work stations with a desk and a chair. The desk has a laptop on it and a curtain besides. It also has a blue bed, a blue sofa and a kitchen counter with chairs close to it.

Revision based Prompting

Context: The white chair is near the table.

Instruction: Generate question answering pairs based on the caption.

Response: Question: What color is the chair near the table?

Answer: The chair near the table is white.

Figure 2: 3D-language data generation pipelines.

Примеры результатов



	B-1	B-2	B-3	B-4	METEOR	ROUHE-L	CIDER	EM
VoteNet+MCAN*	28.0	16.7	10.8	6.2	11.4	29.8	54.7	17.3
ScanRefer+MCAN*	26.9	16.6	11.6	7.9	11.5	30	55.4	18.6
ScanQA*	30.2	20.4	15.1	10.1	13.1	33.3	64.9	21.0
LLaVA(zero-shot)	7.1	2.6	0.9	0.3	10.5	12.3	5.7	0.0
flamingo-SingleImage	23.8	14.5	9.2	8.5	10.7	29.6	52	16.9
flamingo-MultiView	25.6	15.2	9.2	8.4	11.3	31.1	55	18.8
BLIP2-flant5-SingleImage	28.6	15.1	9.0	5.1	10.6	25.8	42.6	13.3
BLIP2-flant5-MultiView	29.7	16.2	9.8	5.9	11.3	26.6	45.7	13.6
3D-LLM (flamingo)	30.3	17.8	12.0	7.2	12.2	32.3	59.2	20.4
3D-LLM (BLIP2-opt)	35.9	22.5	16.0	9.4	13.8	34.0	63.8	19.3
3D-LLM (BLIP2-flant5)	39.3	25.2	18.4	12.0	14.5	35.7	69.4	20.5

Table 1: Experimental results on ScanQA validation set. * Means the models use explicit object representations. B-1, B-2, B-3, B-4 denote BLEU-1, BLEU-2, BLEU-3, BLEU-4 respectively. Our model outperforms all baseline models for all evaluation metrics except for the EM metric.

	BLEU-1	BLEU-4	METEOR	ROUHE-L	CIDER	EM
SingleImage+MCAN	16.5	0.0	8.4	21.5	38.6	15.8
VoteNet+MCAN*	29.5	6.0	12.0	30.9	58.2	19.7
ScanRefer+MCAN*	27.9	7.5	11.9	30.7	57.4	20.6
ScanQA*	31.6	12.0	13.5	34.3	67.3	23.5
3D-LLM (flamingo)	32.6	8.4	13.5	34.8	65.6	23.2
3D-LLM (BLIP2-opt)	37.3	10.7	14.3	34.5	67.1	19.1
3D-LLM (BLIP2-flant5)	38.3	11.6	14.9	35.3	69.6	19.1

Table 2: Experimental results on ScanQA test set. * Means the models use explicit object representations. Our model outperforms all baseline models for most of the evaluation metrics.

- Тогдашние 2D модели в zero-shot дают низкие результаты для 3д задач
- LLM даёт буст в задаче 3D QA
- Размер модели имеет значение
- 3D признаки важны

SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model



A Relative Relation

- Q Can you confirm if ④ has a greater width compared to ⑥?
- A In fact, ④ might be narrower than ⑥.
- Q Which of these four, ③, ④, ⑤, ⑥ is the tallest?
- A ⑥ is the tallest.
- Q From the viewer's perspective, is the position of ⑥ closer than that of ③?
- A No, ③ is closer.

B Metric Measurement

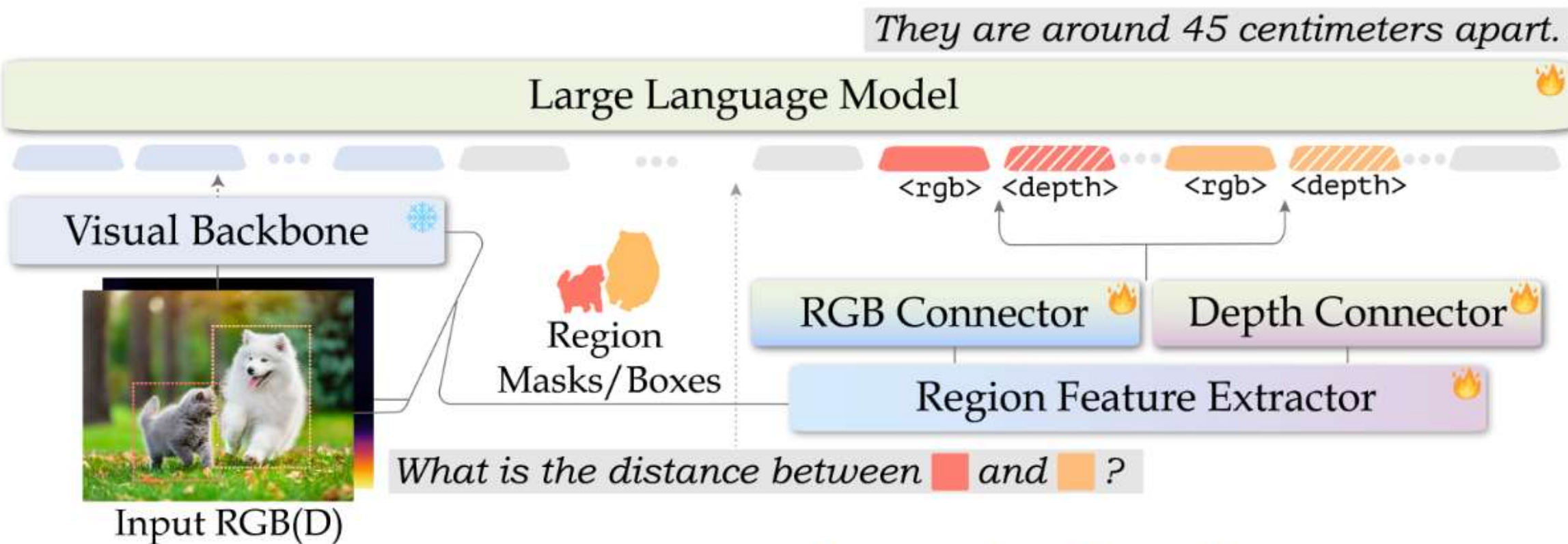
- Q If you are at ③, where will you find ⑦?
- A ⑦ is at 1 o'clock from ③.
- Q What is the height of ④?
- A ④ is 1.38 meters tall.
- Q How far apart are ② and ③ horizontally?
- A They are 15.33 meters apart horizontally.



C Spatial Reasoning

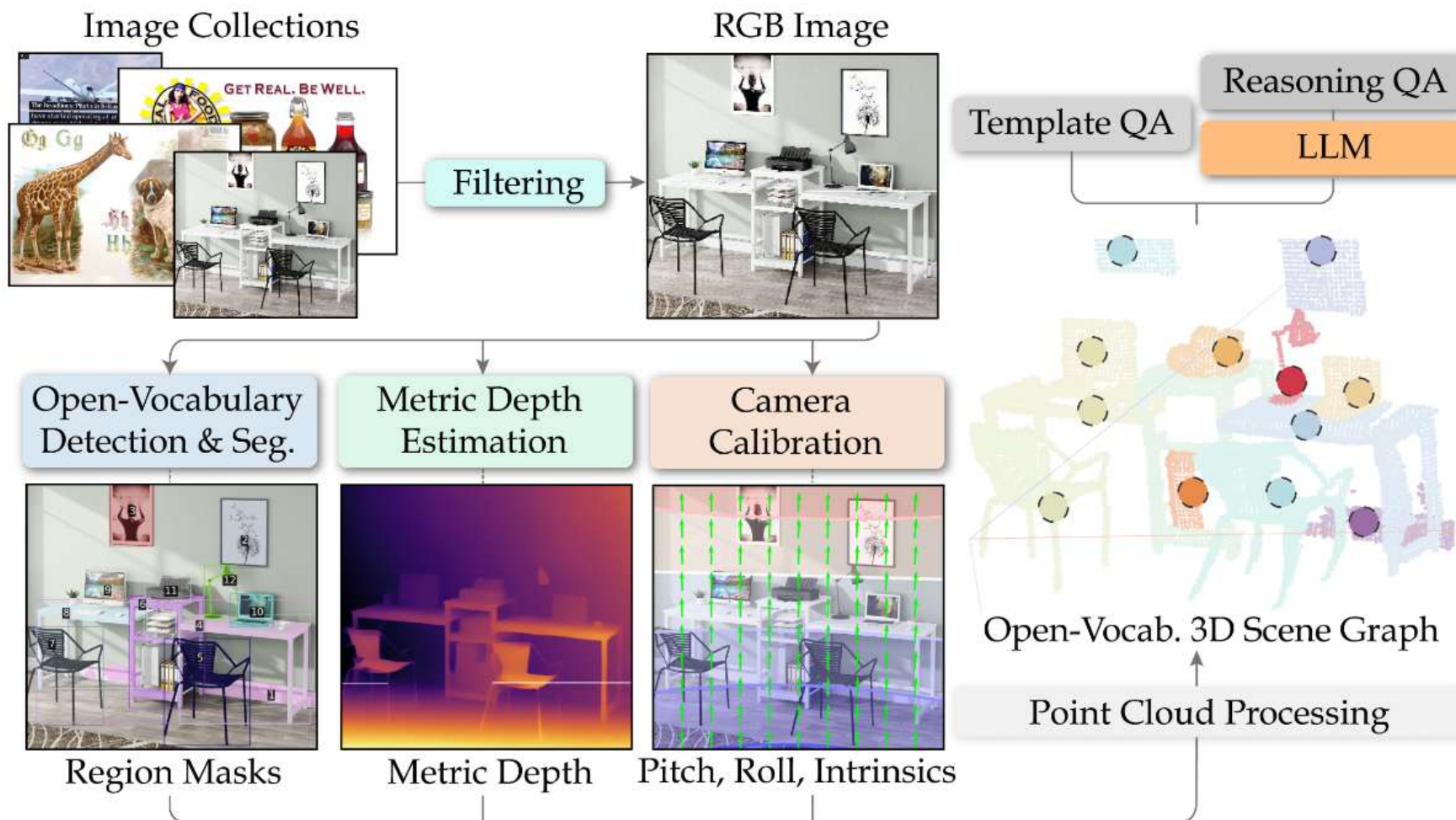
- Q Can you try to estimate how many floors are in ①? Think step by step.
- A The height of ① is 204.54 feet. Assuming each floor is about 10 feet high, the total number of floors would be 20.454. Since you can't have a fraction of a floor, the total number of floors would be approximately 20.
- Q I'm planning to build a bridge from ② to ①, is it possible?
- A The distance between ② and ① is 14.82 meters, so it is possible to build a bridge between them.
- Q If you are riding a motorcycle with 36 inches wide, do you think you can pass through the area between ③ and ④?
- A The distance between ③ and ④ is 38.95 inches, so yes, you can pass through the area between ③ and ④ since the motorcycle is narrower than the distance between them.

SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model



An architecture overview of Spatial RGPT. ❄️ 🔥 denotes frozen/trainable parameters.

SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model

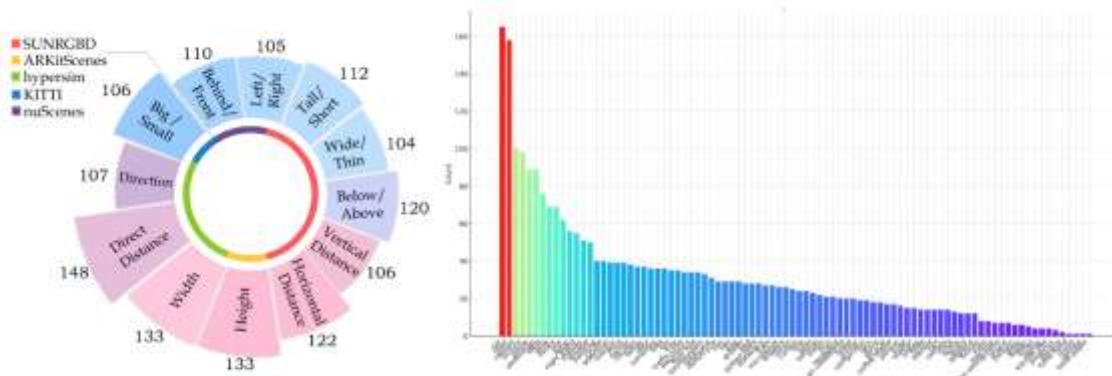


The figures shows our automatic data curation pipeline using from single images.

- Open Spatial Dataset
- Полностью автоматические делается по RGB изображениям
- 8.7M spatial concepts grounded in 5M unique regions from 1M images from OpenImages
- Filtering + open-vocabulary image tagging model + Grounding dino for bbox + SAM-HQ for segmentation
- Metric3Dv2 for depth + WildCamera for camera + PerspectiveFields для перевода



SpatialRGPT: Grounded Spatial Reasoning in Vision Language Model



Our SpatialRGPT-Bench includes 657 qualitative and 749 quantitative VQA pairs, covering 88 distinct classes.



How far is ① from ② horizontally?

① and ② are 31.21 feet apart horizontally.



Could you tell me the vertical size of ①?

① is 3.67 feet tall.



How tall is ① in terms of height?

① is 4.65 inches in height.



Does ① have a larger size compared to ②?

Correct, ① is larger in size than ②.



Does ① have a lesser width compared to ②?

No, ① is not thinner than ②.



Is the position of ① less distant than ②?

No.

Samples from our SpatialRGPT-Bench.

Method	Below/ Above	Left/ Right	Big/ Small	Tall/ Short	Wide/ Thin	Behind/ Front	Avg.
GPT-4	64.1	42.8	42.8	61.6	61.6	49.0	57.8
GPT-4V	63.3	46.6	64.1	60.7	68.2	45.4	58.1
LLaVA-v1.6-34B	44.1	45.7	36.7	53.5	37.5	45.4	43.9
GPT-4V+SoM	75.0	55.2	42.4	54.4	49.0	47.2	54.3
LLaVA-v1.6-34B+SoM	44.1	40.0	33.9	47.3	41.3	46.3	42.3
Kosmos-2	28.3	15.2	4.71	26.7	12.5	12.7	17.0
RegionVILA	30.8	47.6	35.8	44.6	35.5	49.0	40.4
SpatialRGPT	99.1	99.0	79.2	89.2	83.6	87.2	89.8
SpatialRGPT-Depth	99.1	99.0	80.1	91.9	87.5	91.8	91.7

SpatialRGPT-Bench Qualitative Results: Numbers represent success rates in percentage (↑).

Method	Direct Distance		Horizontal Distance		Vertical Distance		Width		Height		Direction	
GPT-4	21.6	1.29	11.5	2.08	33.0	0.65	52.3	0.52	48.1	1.40	34.6	83.7°
GPT-4V	29.7	0.92	25.4	2.75	33.0	0.48	51.1	0.37	68.4	1.57	43.9	69.9°
LLaVA-v1.6-34B	24.3	0.76	24.5	1.59	30.1	0.62	30.8	0.40	42.8	1.96	33.6	78.2°
GPT-4V+SoM	25.7	1.02	22.1	2.36	33.9	0.64	45.8	0.70	62.4	1.08	54.2	55.5°
LLaVA-v1.6-34B+SoM	12.8	1.15	20.4	1.79	11.3	0.95	9.02	0.91	7.52	3.11	12.8	33.3°
Kosmos-2	4.05	>10	4.91	>10	1.89	2.26	3.01	5.42	1.50	3.82	1.86	104°
RegionVILA	22.3	1.30	24.6	3.26	17.9	>10	36.8	>10	49.6	1.61	35.5	79.8°
SpatialRGPT	35.1	0.35	59.0	0.27	53.8	0.27	51.9	0.31	54.9	0.63	95.3	17.1°
SpatialRGPT-Depth	41.2	0.33	65.6	0.25	51.9	0.27	49.6	0.31	57.9	0.61	95.3	15.4°

SpatialRGPT-Bench Quantitative Results: Numbers represent success rates within $\pm 25\%$ of the ground-truth in percentage (↑) and absolute relative error in metric scale (↓).

LLaVA-3D



(a)

3D Dense Captioning



The bed is in the back left corner of the room. It is to the right of the window on the left wall.

3D Visual Grounding

Q: "A door sits partially open to the right of a wooden cabinet, to the doors left is another door with a window." Which object best matches the given description? Please provide its coordinates.
A: [1.16, 0.95, 0.48, 2.32, 1.10, 0.95].

3D Question Answering

Q: The related object is located at [-0.826, 1.552, 1.153]. What is the region role of this object?
A: The shelf is placed above the desk, making it convenient to access books and study materials.

Image Captioning

The image shows a neatly made twin bed with a wooden frame and a mattress covered in a dark blue fabric. The bed is situated in a room with a window that is partially open, allowing natural light to enter. The window has a wooden frame and a white curtain that is partially drawn, revealing a glimpse of the outside.

VQA

Q: What is the color of the bed?
A: Blue.

2D Visual Grounding

Q: Provide the bounding box coordinate of the bed.
A: [0.12, 0.41, 0.23, 0.51]

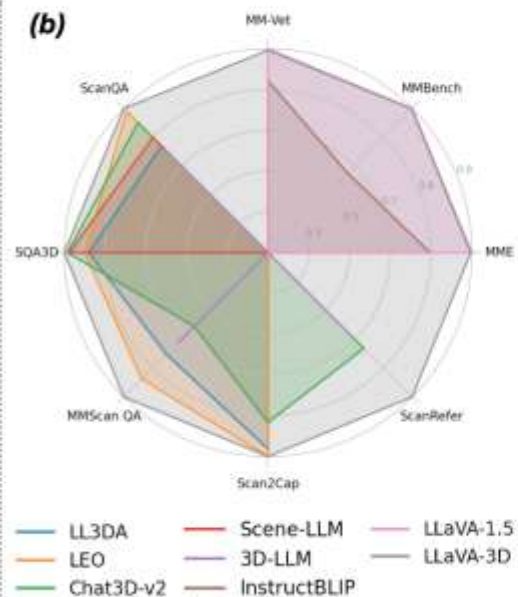
region-level VQA

Q: Is there anything on the bed in the image?
A: No.

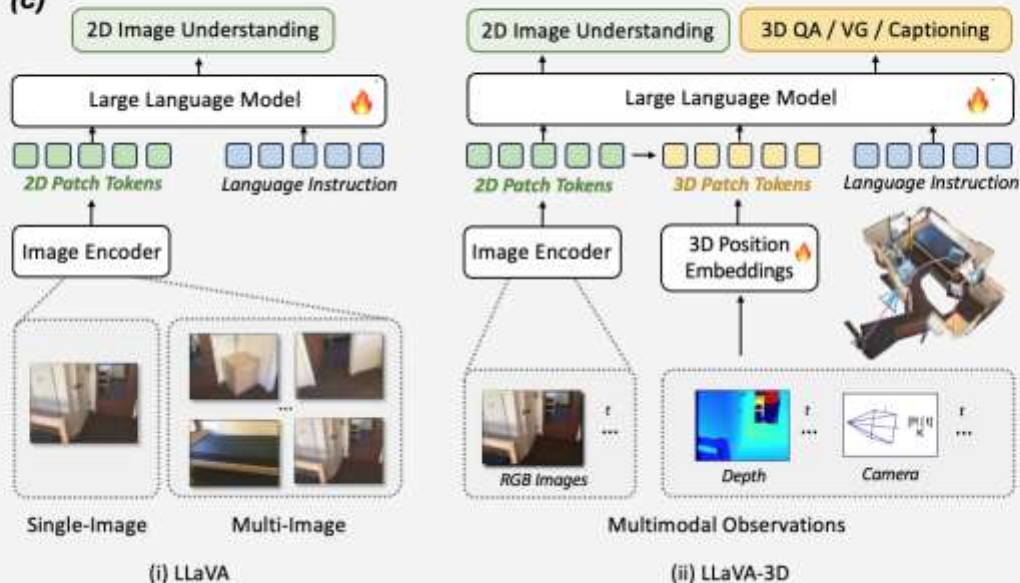
OCR

Multi-Round Conversation

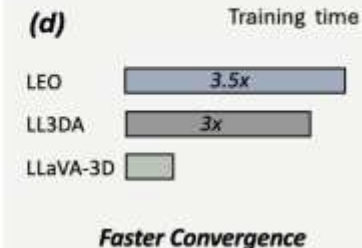
(b)



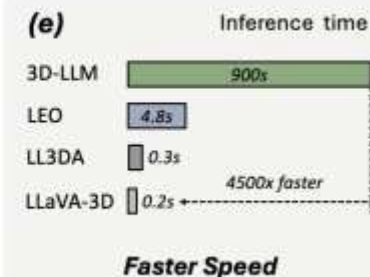
(c)

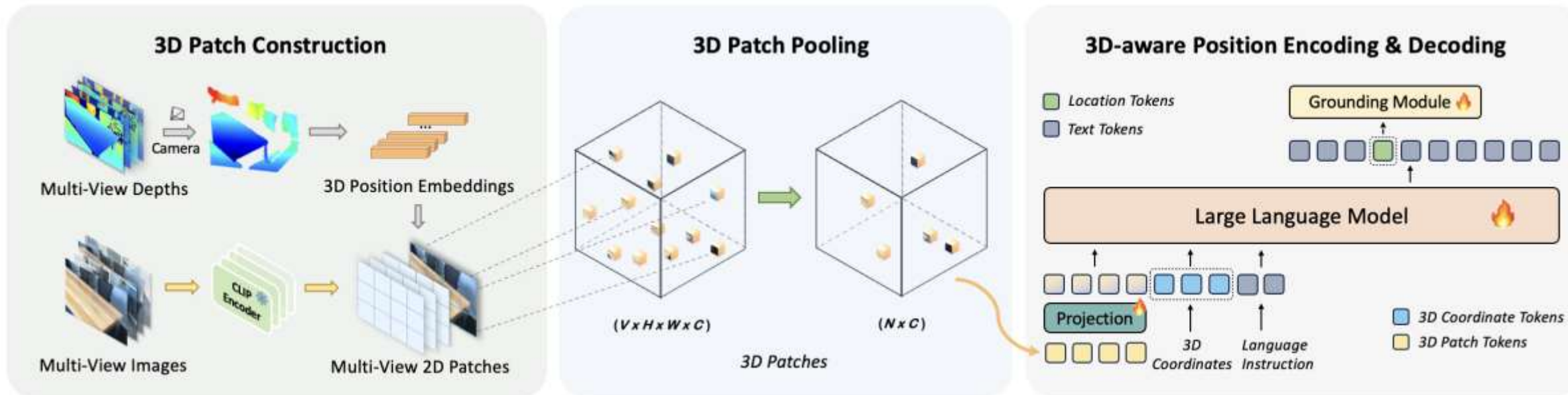


(d)



(e)





- 3D Positional Embeddings учатся через 2х уровневую MLP
- Добавляются к признакам CLIP
- Применяется pooling для сжатия
- Можно использовать просто voxel pooling, можно более умный (но спорный)
- Для запросов с 3D координатами учим простой токенизатор (2-х уровневую MLP)
- Токены-ответы координат декодируем в 3D bbox



Датасет и результаты

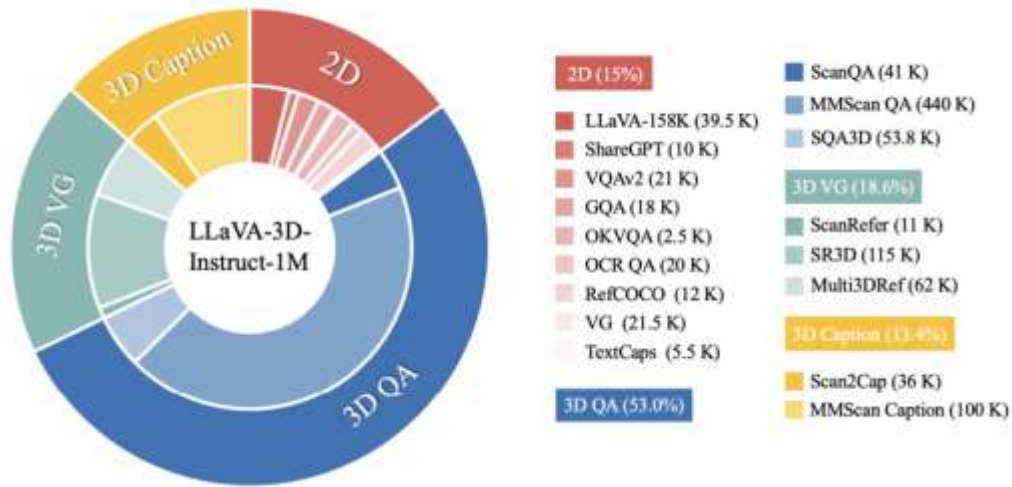


Figure 3. **LLaVA-3D-Instruct-1M**. The hybrid 2D and 3D Dataset Collection. Left: Distribution of data across categories, with the outer circle representing all categories and the inner circle illustrating data subset distribution. Right: Detailed dataset quantities.

	ScanQA (val)					SQA3D (test)
	C	B-4	M	R	EM@1	EM@1
<i>Task-specific models</i>						
Scan2Cap [11]	-	-	-	-	-	41.0 [†]
ScanRefer+MCAN [49]	55.4	7.9	11.5	30.0	18.6	-
ClipBERT [25]	-	-	-	-	-	43.3
ScanQA [3]	64.9	10.1	13.1	33.3	21.1	47.2
3D-VisTA [53]	69.6	10.4	13.9	35.7	22.4	48.5
<i>Task-specific fine-tuned 3D LLMs</i>						
3D-LLM (FlanT5) [16]	69.4	12.0	14.5	35.7	20.5	-
LL3DA [35]	76.8	13.5	15.9	37.3	-	-
Chat-3D v2 [17]	87.6	14.0	-	-	-	54.7
LEO [18]	101.4	13.2	20.0	49.2	24.5 (47.6)	50.0 (52.4)
Scene-LLM [15]	80	12.0	16.6	40.0	27.2	54.2
<i>Zero-shot 2D LLMs</i>						
VideoChat2 [30]	49.2	9.6	9.5	28.2	19.2	37.3
LLaVA-NeXT-Video [26]	46.2	9.8	9.1	27.8	18.7	34.2
GPT-4V	59.6	-	13.5	33.4	-	-
Gemini	68.3	-	11.3	35.4	-	-
Claude	57.7	-	10.0	29.3	-	-
LLaVA-3D	91.7	14.5	20.7	50.1	27.0 (45.0)	55.6 (57.6)

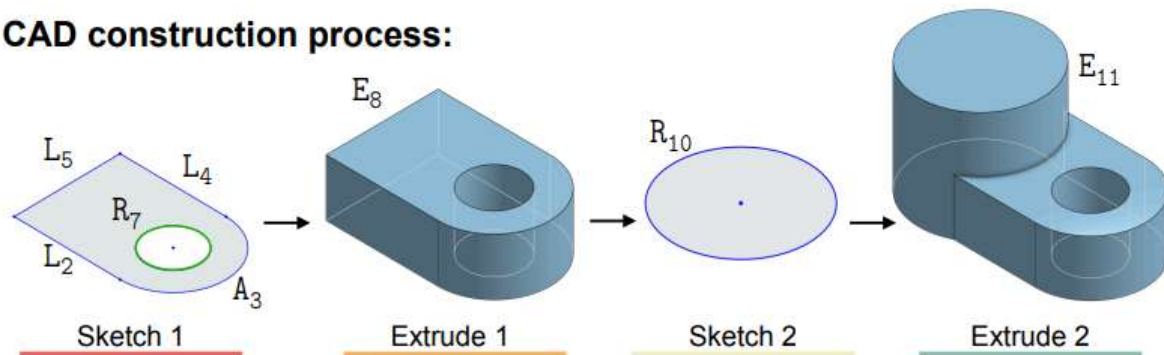


Работа с САД-моделями

CAD-модели



CAD construction process:



Parametrized command sequence:

$\langle \text{SOL} \rangle_1 : \emptyset$ $L_2 : (2, 0)$ $A_3 : (2, 2, \pi, 1)$ $L_4 : (0, 2)$ $L_5 : (0, 0)$ $\langle \text{SOL} \rangle_6 : \emptyset$ $R_7 : (2, 1, 0.5)$	$E_8 : (0, 0, 0, -2, -1, 0, 3,$ $1, 0, \text{New body, One-sided})$ $\langle \text{SOL} \rangle_9 : \emptyset$ $R_{10} : (0, 0, 1.125)$ $E_{11} : (0, 0, 0, -2, 0, 0, 2.25,$ $2, 0, \text{Join, One-sided})$ $\langle \text{EOS} \rangle_{12} : \emptyset$
---	--

Commands	Parameters
$\langle \text{SOL} \rangle$	\emptyset
L (Line)	x, y : line end-point
A (Arc)	x, y : arc end-point α : sweep angle f : counter-clockwise flag
R (Circle)	x, y : center r : radius
E (Extrude)	θ, ϕ, γ : sketch plane orientation p_x, p_y, p_z : sketch plane origin s : scale of associated sketch profile e_1, e_2 : extrude distances toward both sides b : boolean type, u : extrude type
$\langle \text{EOS} \rangle$	\emptyset

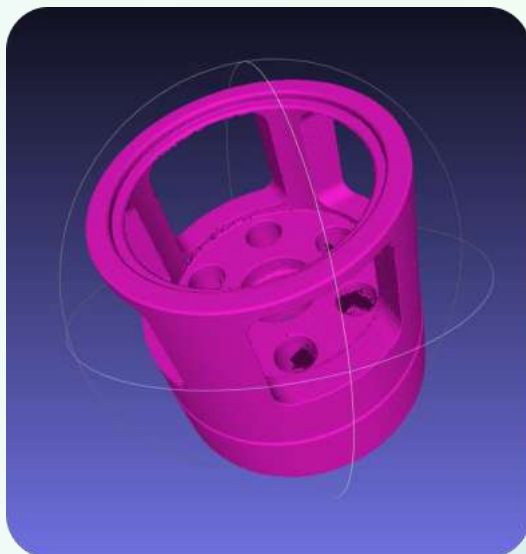
- Loop - одна замкнутая кривая
- Sketch profile – набор loop в одной плоскости
- Extrude - "вытягивание" в 3D с каким-то взаимодействием с предыдущими 3D формами

Задача реверс-инжиниринга для mechanical design

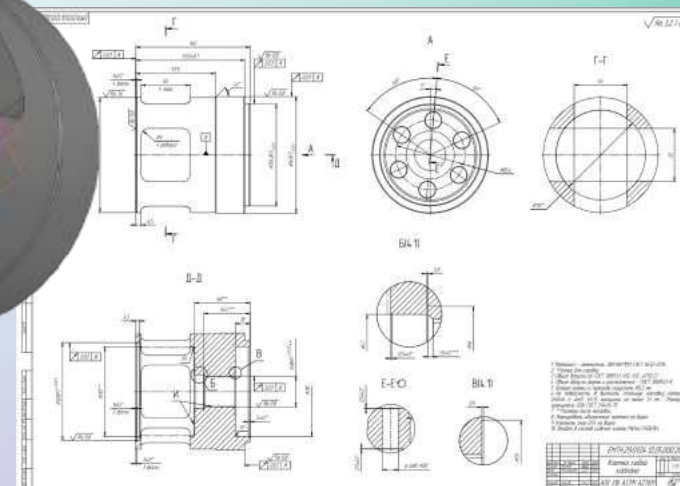
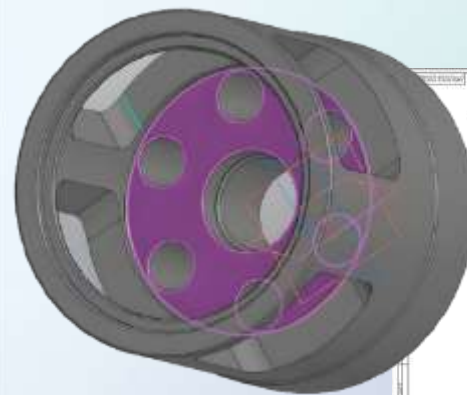


Суть – построение полноценной конструктивной модели (CAD) выбранной детали в инженерном ПО, с целью производства точной копии детали.

Реальный объект и его 3D скан



Цифровая модель и чертёж



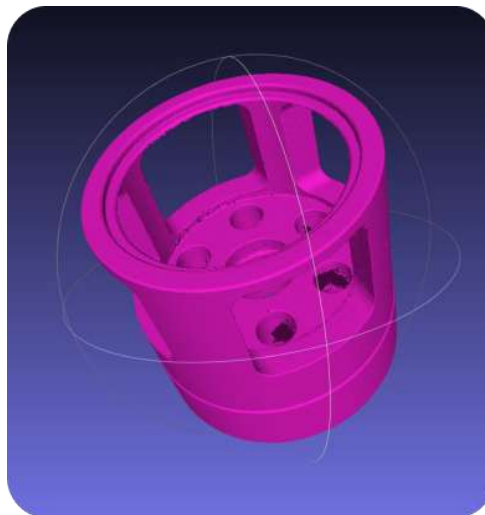
Результат реверс-инжиниринга и 3Д реконструкция



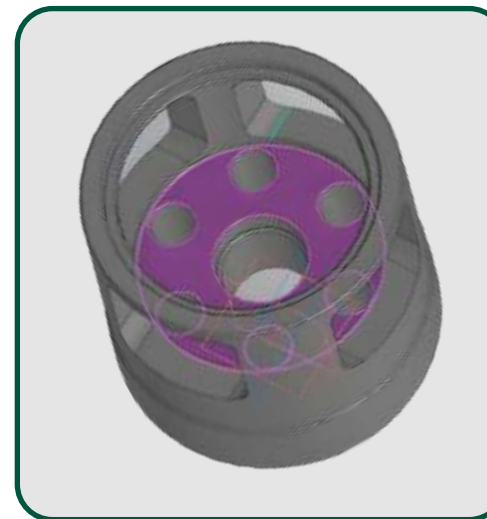
Задача реверс-инжиниринга похожа на задачу 3D реконструкции (3D computer vision), но итоговая модель представлена не в привычном mesh, а в параметрическом граничном представлении B-rep плюс пошаговая история получения этого представления из конструктивных операций



Фотография
детали



Сетка



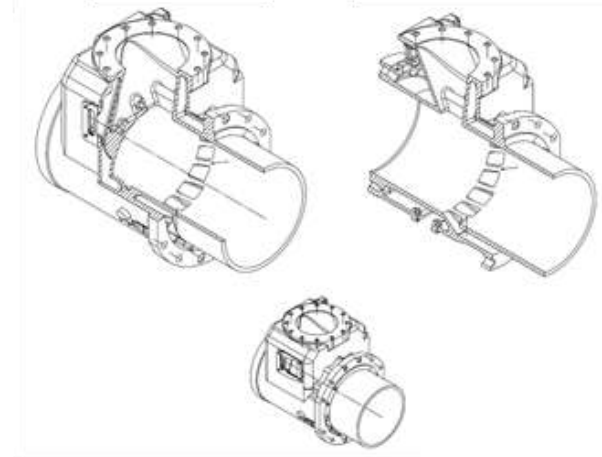
B-rep с историей
построения

Важно – (а) нужна высокая точность B-rep (б) операции конструктивного построения могут отличаться между конкретными инженерными ПО

Реальный процесс реверс-инжиниринга



Реальное
устройство из
множества
деталей



3D модели всех
деталей и сборка
в Компас3Д

- Если деталь в устройстве, то устройство последовательно разбираем на части, процесс документируем
- На деталь наклеиваются калибровочные метки и она сканируется ручным сканером в 3D
- Деталь фотографируется, проводятся ручные измерения отдельных элементов (фаски, отверстия, резьба и т.д.)
- Скан (меш) модели загружается в ПО. Инженер вручную конструирует деталь, чтобы результат совпадал со сканом
- Менеджер валидирует результат

Деталь может быть сильно повреждена



До



После





Разделы предметной области:

1. Бенчмарки.

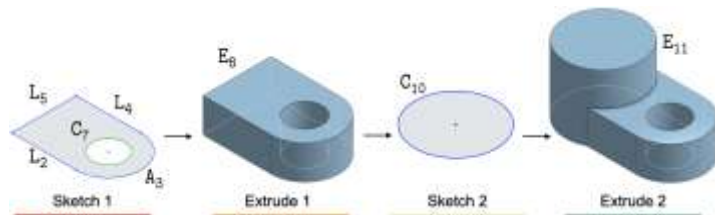
- Данные, на которых оценивается качество решения задачи.
- Метрики, которые используются для оценки качества решения задачи.

2. Методы решения задачи

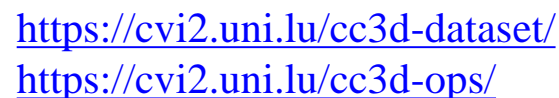
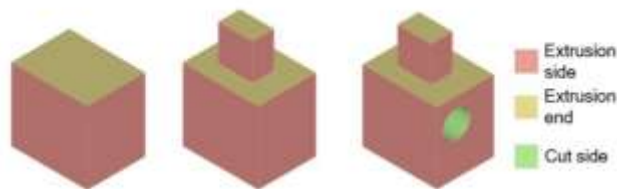
- End-2-end подходы
- Композиция из методов решения отдельных подзадач

3. Данные для обучения

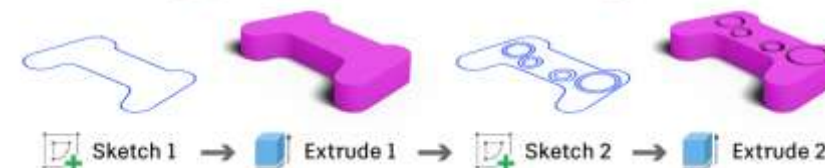
- Датасеты из реальных примеров (часто общие с данными для бенчмарков)
- Методы генерации синтетических данных



- 178,238 моделей из OnShape системы, взято подмножество ABC датасета
- 8046 моделей в бенчмарк



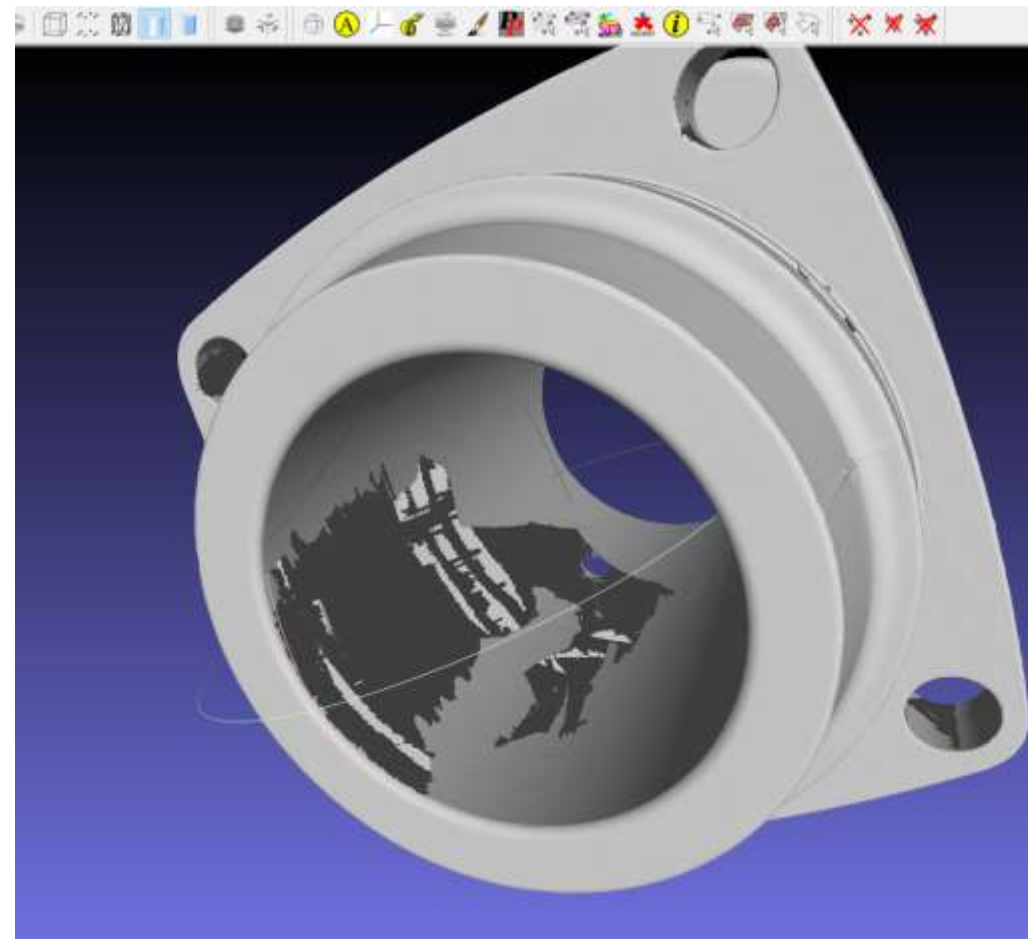
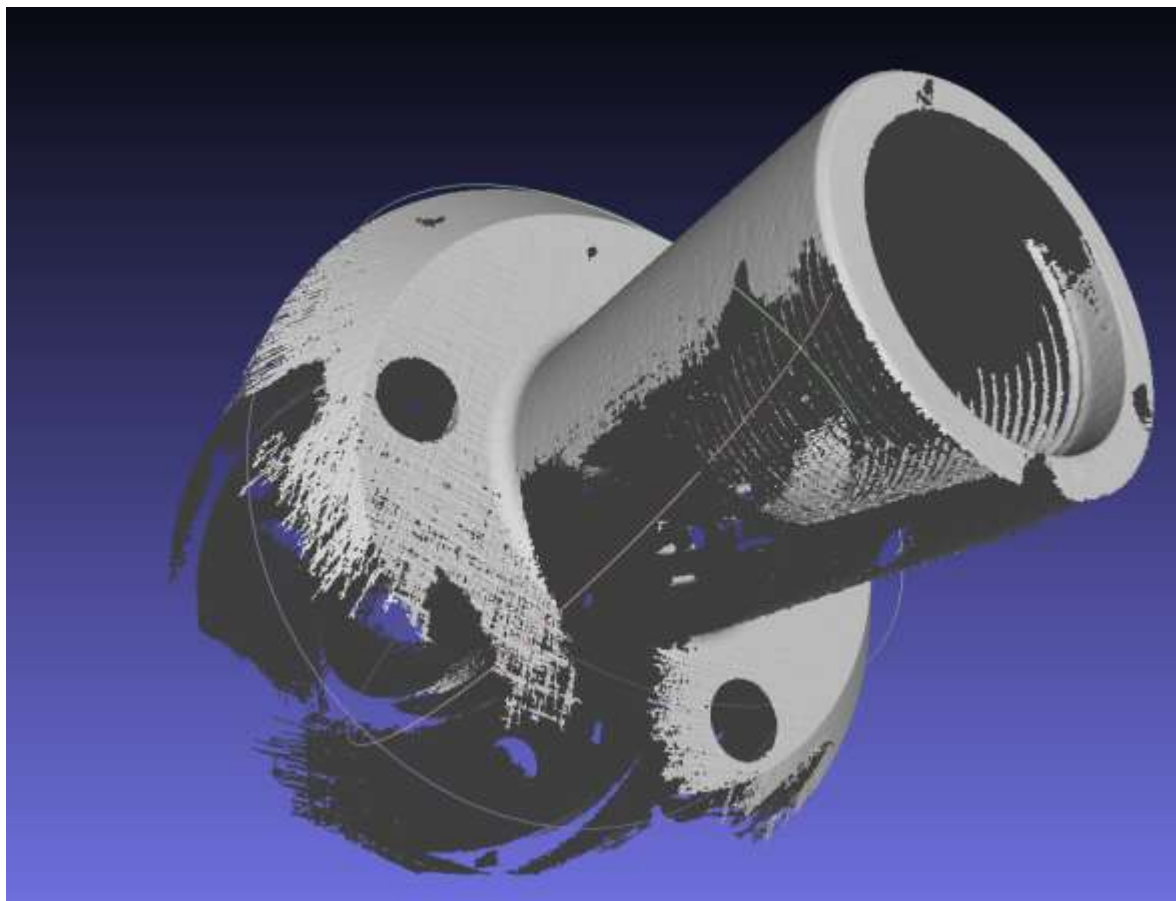
- 37k+ моделей
- 2973 в бенчмарк



1725 моделей в бенчмарк

- Десятки тысяч моделей
- Простой набор операций
- Синтетические модели без реальных сканов

Особенности реальных сканов



Отсутствуют куски поверхности в скане из-за ограниченности ракурсов сканирования на практике

Особенности реальных сканов

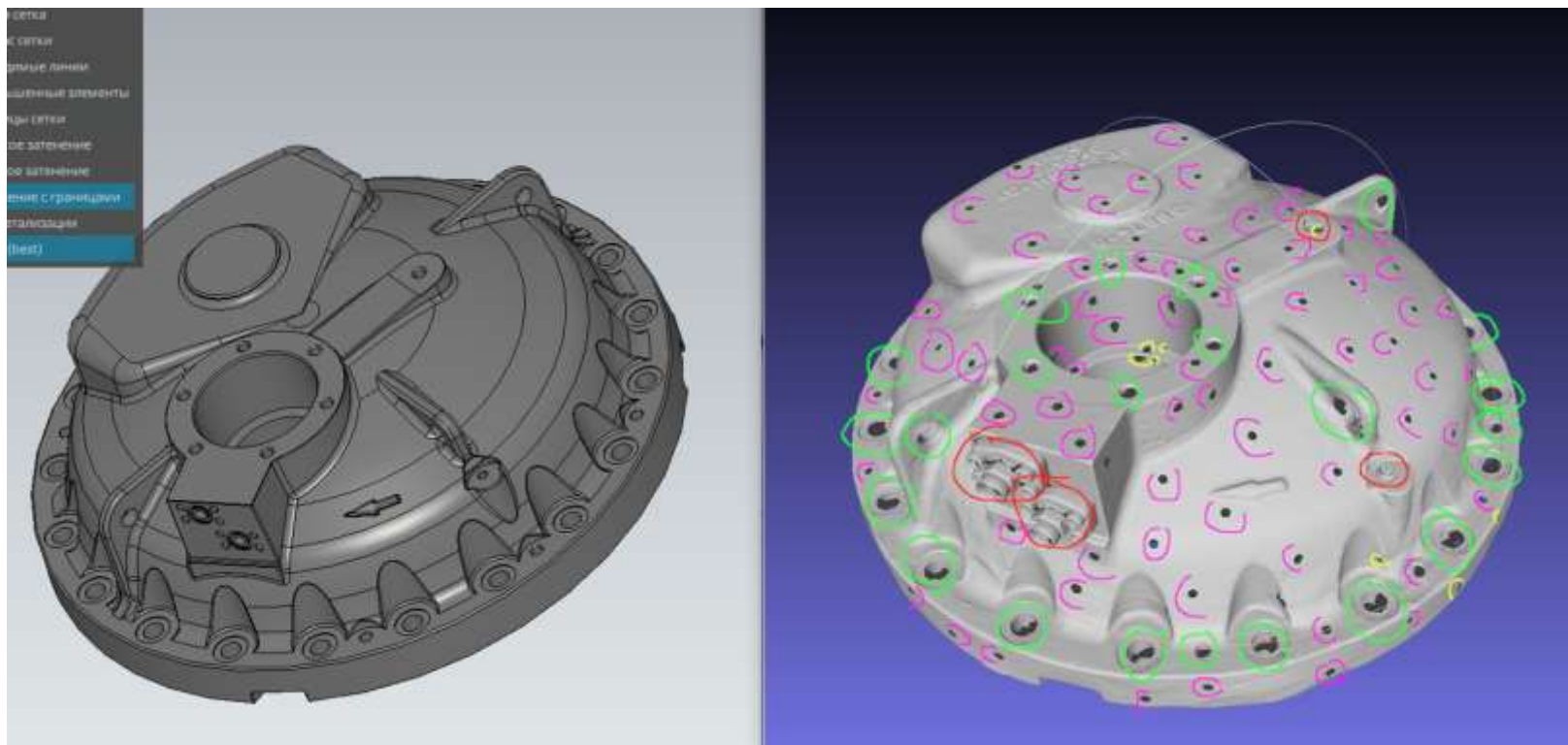


1) ■ случайные дырки в меше
(надо заделать)

2) ■ круглые дырки в меше из-за
меток (надо заделать)

3) ■ недосканированные круглые
отверстия в самой детали которые
выглядят почти как круглые дырки
в меше (нельзя заделывать)

4) ■ Часть меша принадлежит
другой детали



Множество небольших артефактов разной природы

Анализ операций и сложности



Этап 1: Формообразующие операции

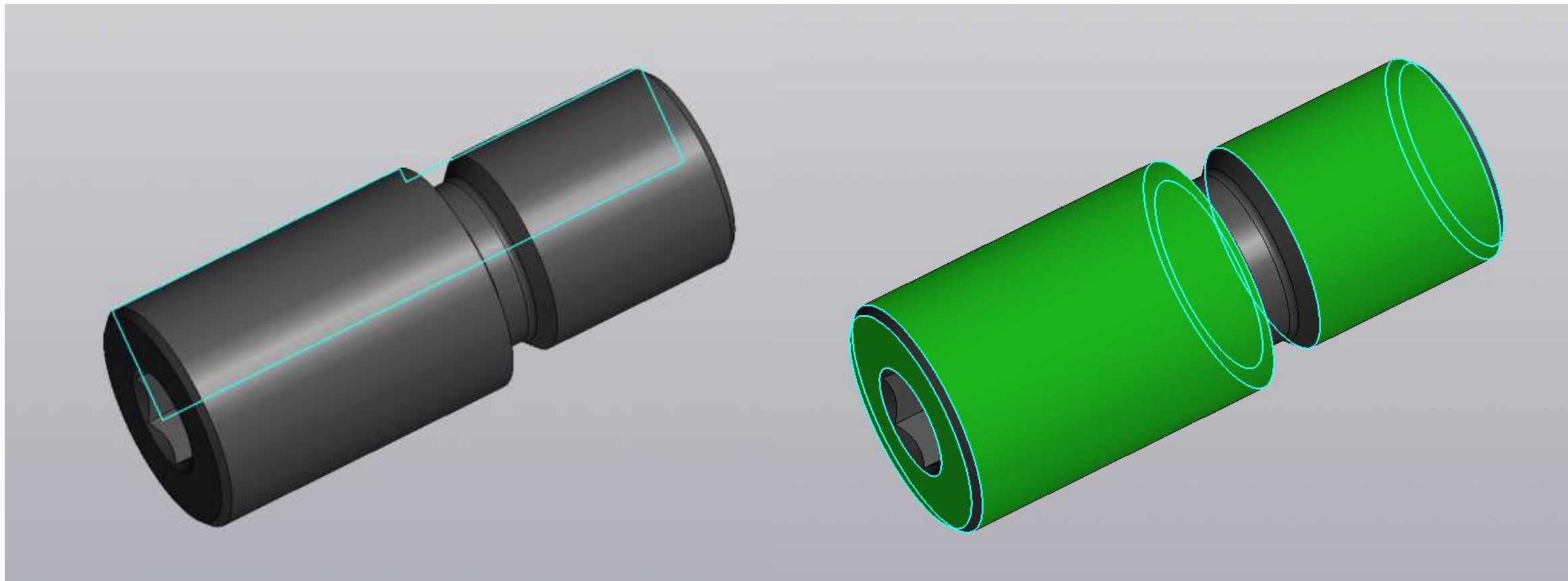
1. **Выдавливание (Extrude)**
2. **Вращение (Revolve)**
3. **Элемент по траектории (Sweep)**
4. **Элемент по сечениям (Loft)**
5. **Булевы операции (Boolean)**
6. Обечайка (Shell)
7. **“Операция без истории”**
8. Отдельные детали по ГОСТу

- **очень часто**
- **довольно часто**
- **редко**
- **один раз**

Этап 2: Дорабатывающие операции

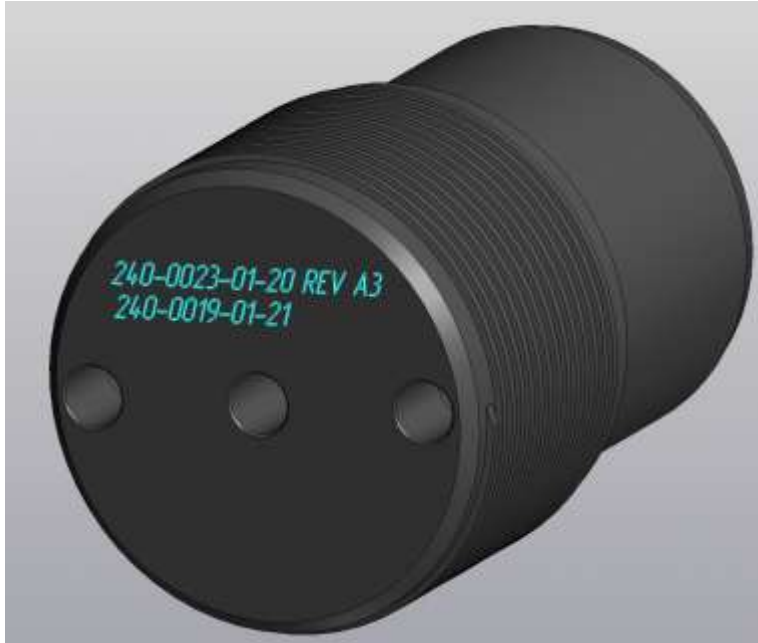
1. **Отверстие (Hole)**
2. **Фаски (Chamfer)**
3. **Скругления (Fillet)**
4. **Массивы (Array, Pattern)**
5. **Надписи / гравировки (Text)**
6. **Резьба**
7. **Прочие детали оформления: “отверстие ГОСТ 14034-74” (Model 30.7), условные обозначения и т. п.**

Пример - Вращение как формообразующая операция

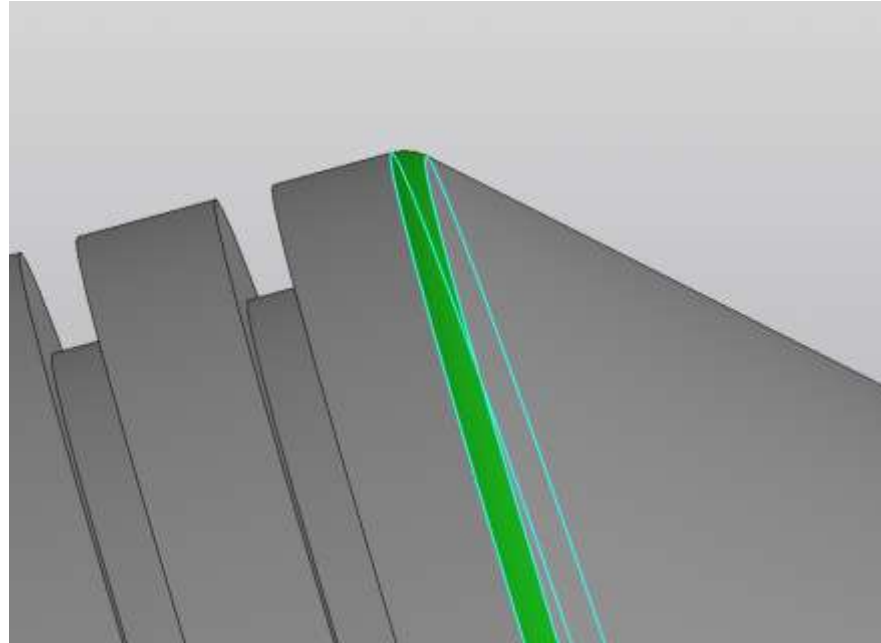


Винт сердечника

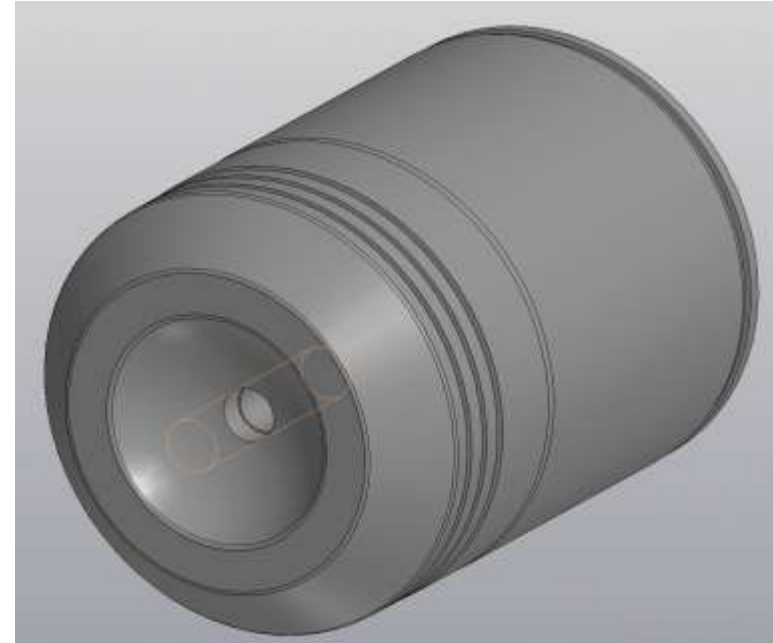
Примеры дорабатывающих операций



Надпись



Скругление

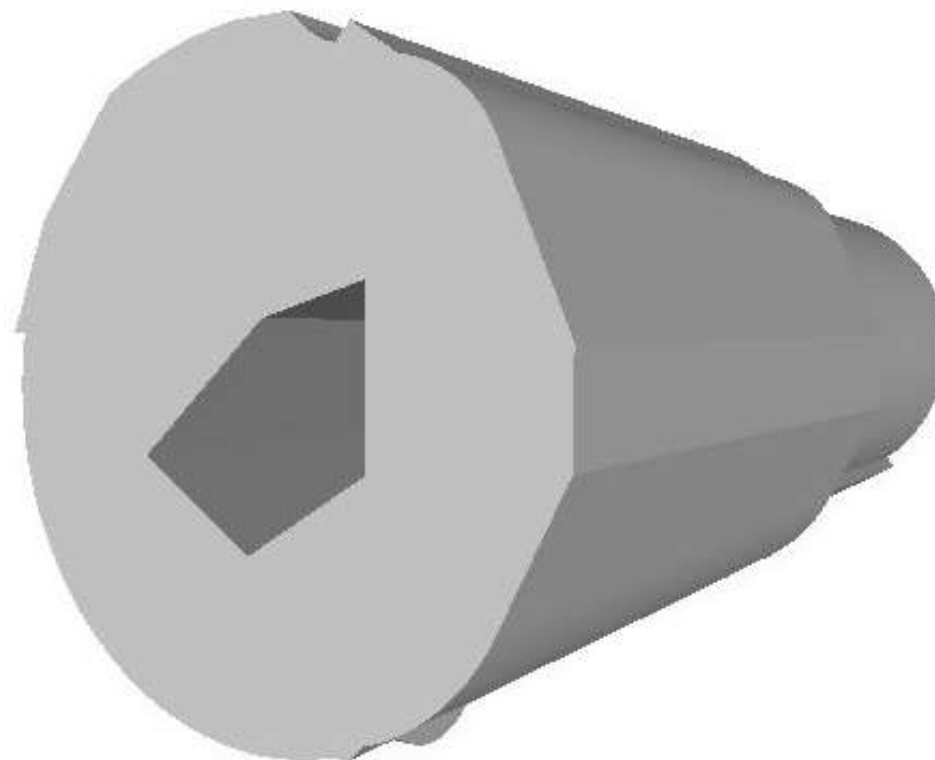
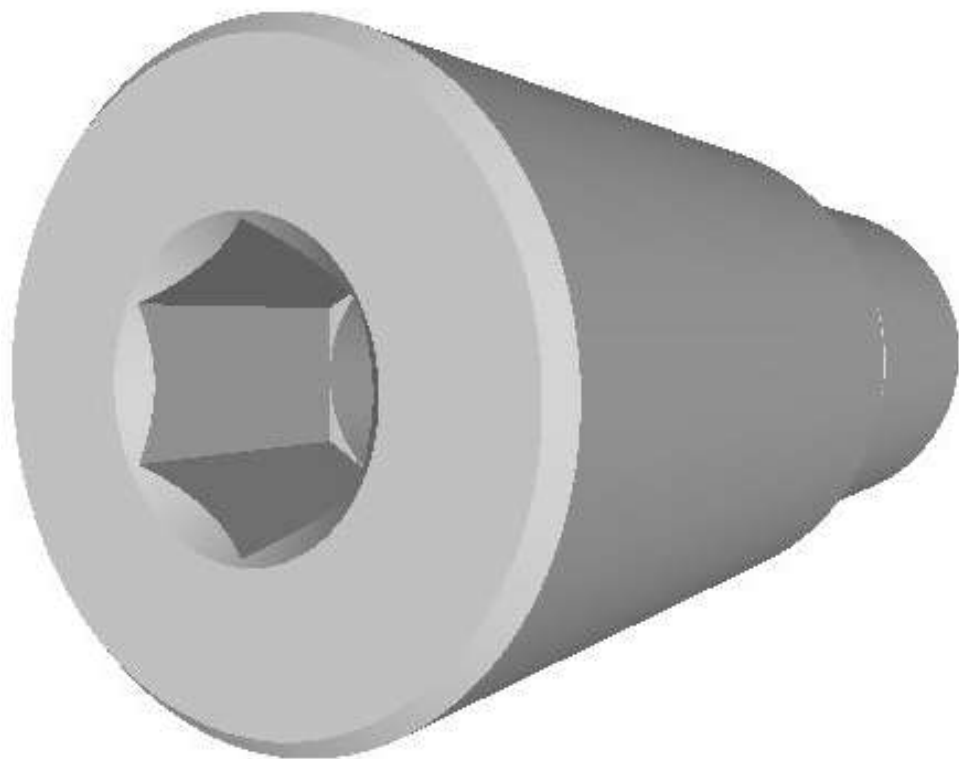


Отверстие



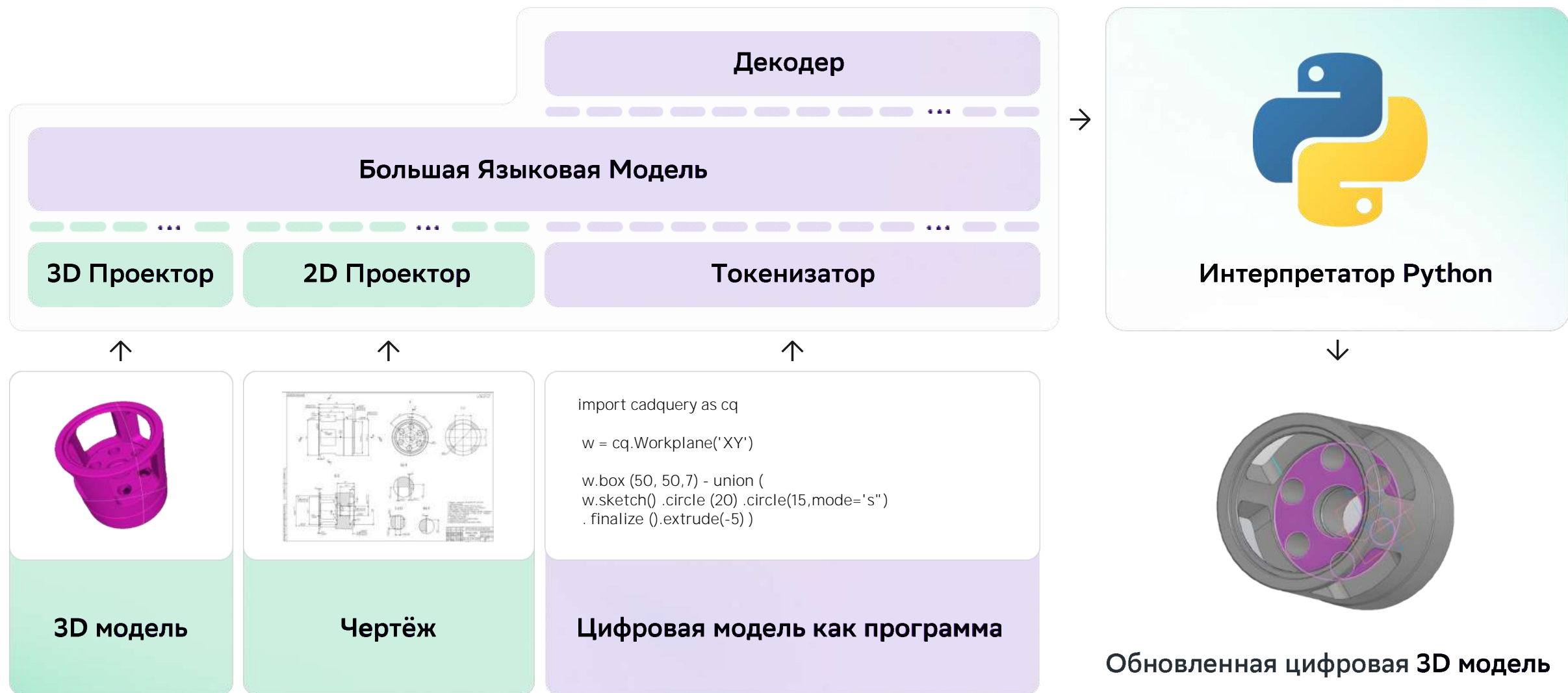
Method	Train Dataset		DeepCAD Test Set				Fusion360 Test Set			
	Name	Size	Mean CD↓	Med. CD↓	IoU↑	IR ↓	Mean CD↓	Med. CD↓	IoU↑	IR↓
DeepCAD [48]	DeepCAD	160k	42.5	9.64	46.7	7.1	330	89.2	39.9	25.2
PrismCAD [25]	DeepCAD	127k	–	4.28	72.1	16.2	–	4.75	65.3	18.0
Point2Cyl [45]	DeepCAD	35k	–	4.27	73.8	3.9	–	4.18	67.5	3.2
HNC-CAD [52]	DeepCAD	125k	–	8.64	65.3	5.6	–	36.8	63.5	7.3
MultiCAD [33]	DeepCAD	160k	–	8.09	–	11.5	–	42.2	–	16.5
TransCAD [12]	DeepCAD	140k	32.3	4.51	65.5	1.1	78.6	33.4	60.2	2.4
CAD-Diffuser [34]	DeepCAD	160k	–	3.02	74.3	1.5	–	3.85	63.2	1.7
CAD-SIGNet [21]	DeepCAD	160k	3.43	0.28	77.6	0.9	7.37	0.48	65.6	1.6
CAD-Recode	DeepCAD	160k	1.98	0.27	80.7	0.0	3.37	0.52	67.6	0.1
CAD-Recode	Ours	1M	0.30	0.16	92.0	0.4	0.35	0.15	87.8	0.5

Table 1. Comparison of CAD reverse engineering methods on DeepCAD and Fusion360 datasets. Our CAD-Recode trained on the 160 k DeepCAD dataset demonstrates an improvement over existing state-of-the-art methods both in terms of geometric fidelity and validity of the generated sketch-extrude sequences. Our procedurally generated dataset provides a significant boost in the prediction quality.



- CD2: $4.5e-3$
- IoU: 0.89

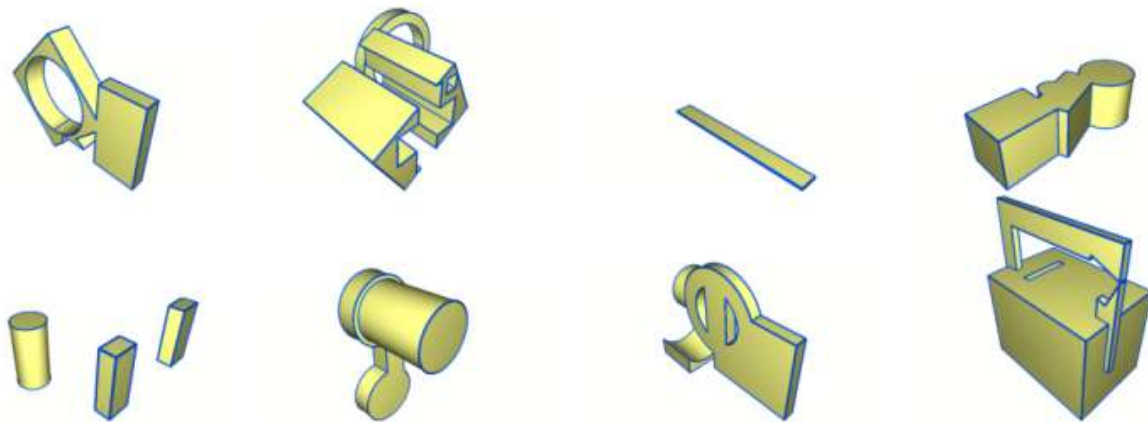
Методы – Мультимодальные LLM



Датасеты для обучения



Рандомизированный подход:



[CAD-Recode: Reverse Engineering CAD Code from Point Clouds](#)

- Рандомизированная генерация примеров на базе операций (profile, extrude)
- Примеры не очень осмысленные

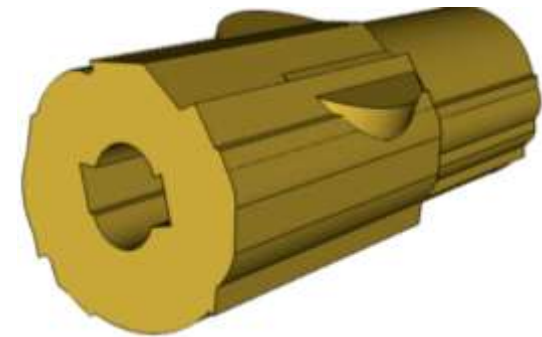
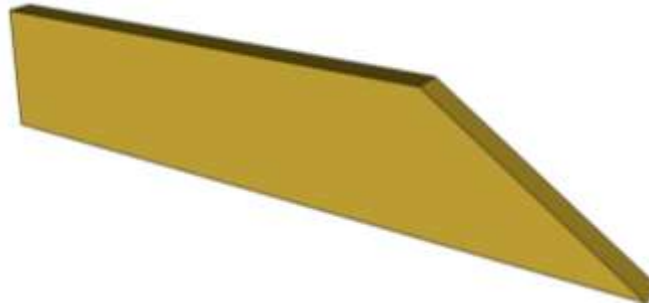
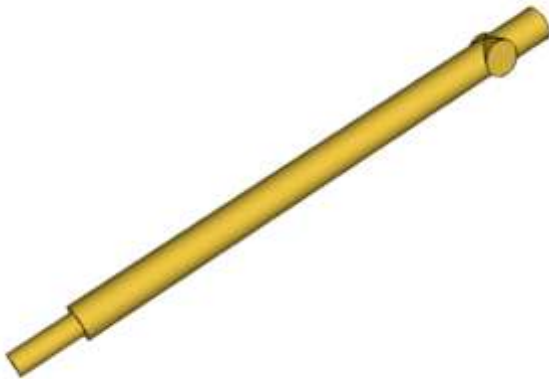
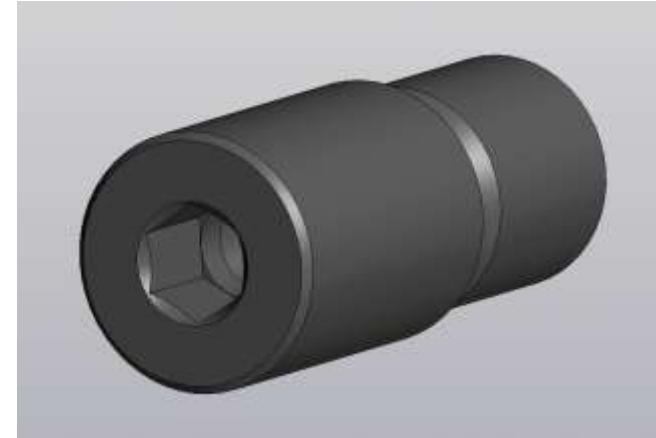
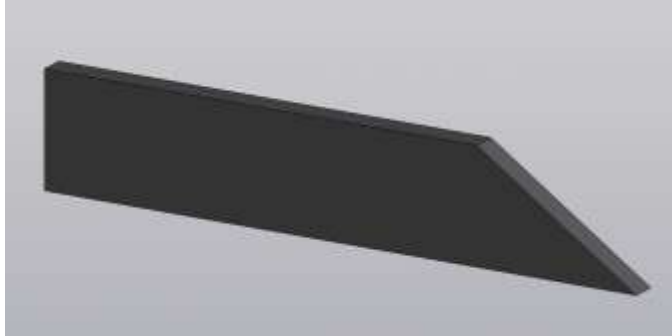
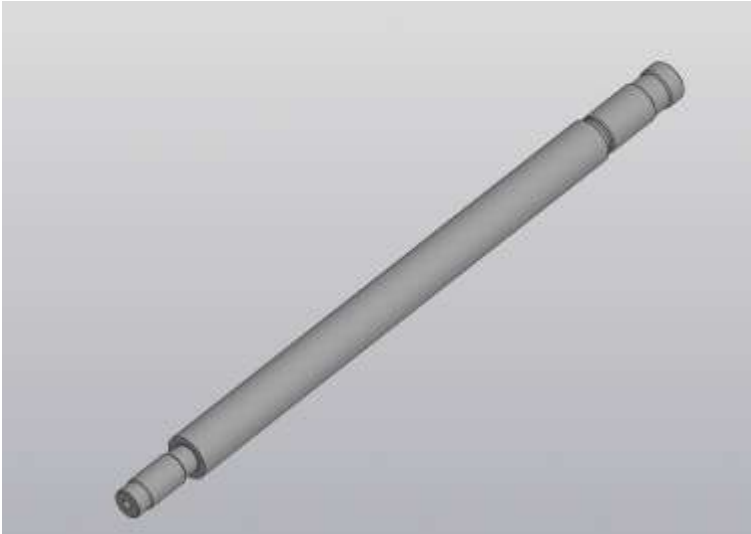
LLM-based:



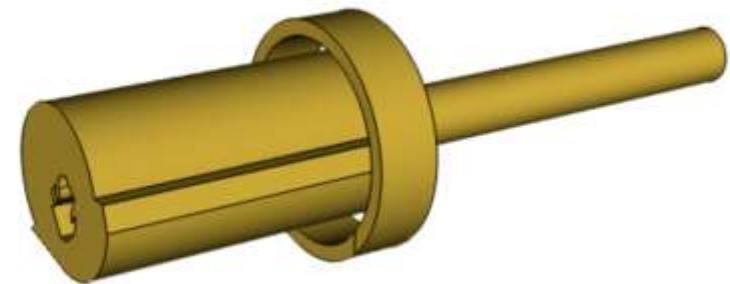
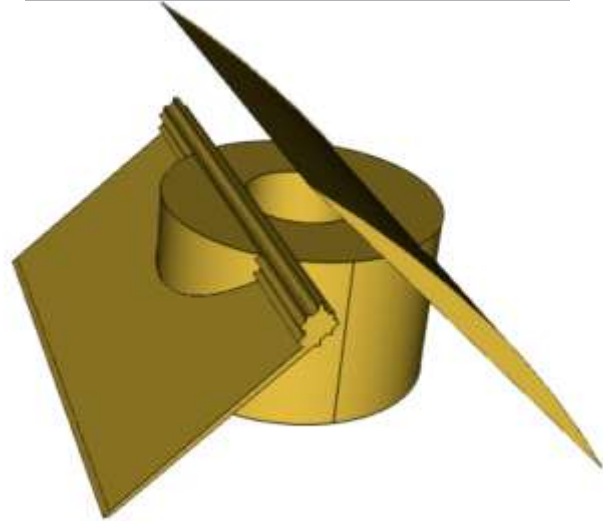
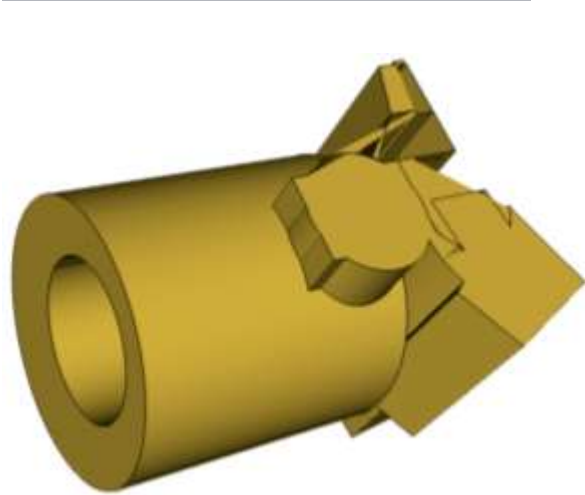
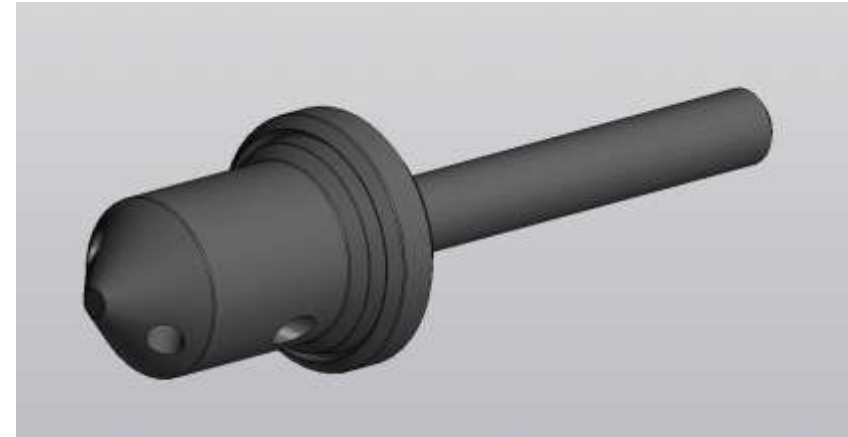
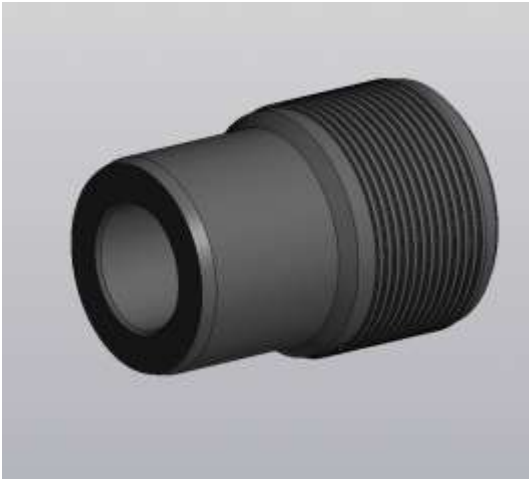
Подход:

- Выделить 10-20 частых типов деталей
- Промтнуть LLM, чтобы по описанию класса деталей, получить параметризацию деталей
- Нагенерировать 1000 моделей для каждого класса
- Использовать хорошие примеры генераторов классов для генерации широкого набора классов деталей

Актуальные примеры работы (Хорошие)



Актуальные примеры работы (Плохие)



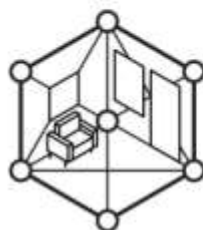
CAD-модели для 3D реконструкции помещений



Scene walkthrough



SceneScript



Neural network



```
1 make_wall,  
2 {id = 0},  
3 {a_x = -0.518},  
4 {a_y = 7.228},  
5 {a_z = 8.0},  
6 {b_x = -0.95},  
7 {b_y = 7.228},  
8 {height = 2.089},  
9 ...  
10 make_door,  
11 {id = 2000},  
12 {wall_id = 0},  
13 {position_x = -1.028},  
14 {position_y = 7.228},  
15 {size_x = 2.533},  
16 {size_y = 1.991},  
17 door_window,  
18 {id = 2001},  
19 {wall_id = 2},  
20 {position_x = -1.348},  
21 {position_z = 1.514},  
22 {size_y = 2.735}
```

Structured scene language



Language to geometry interpreter



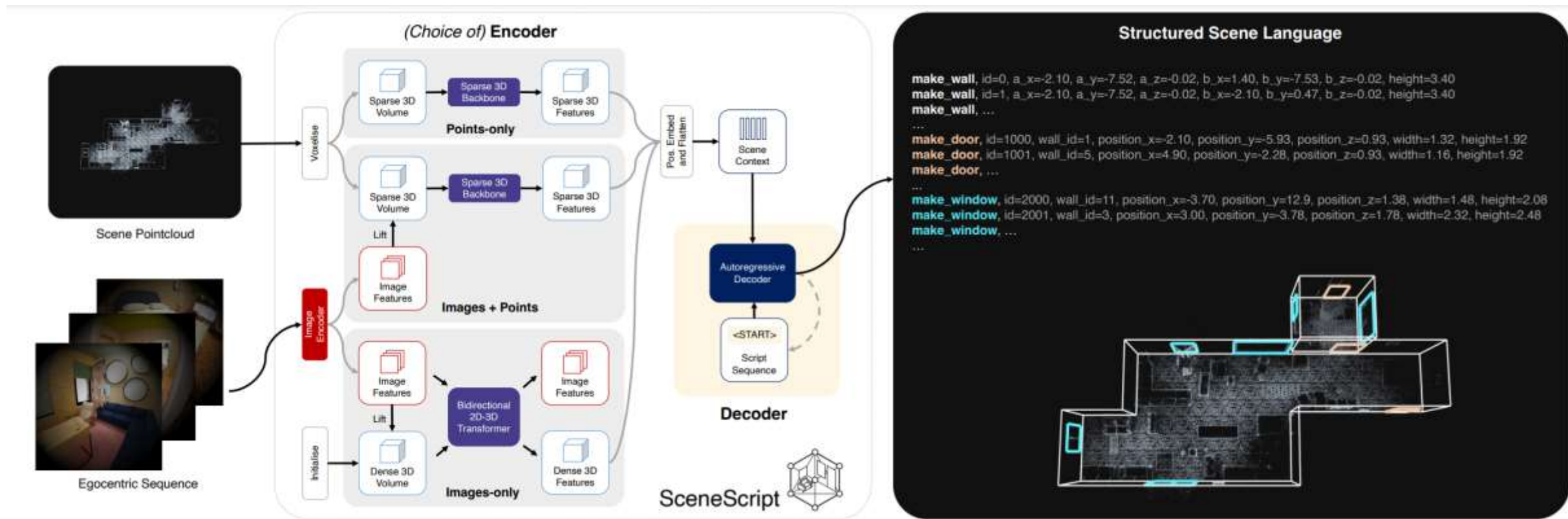
Описание модели через программный скрипт



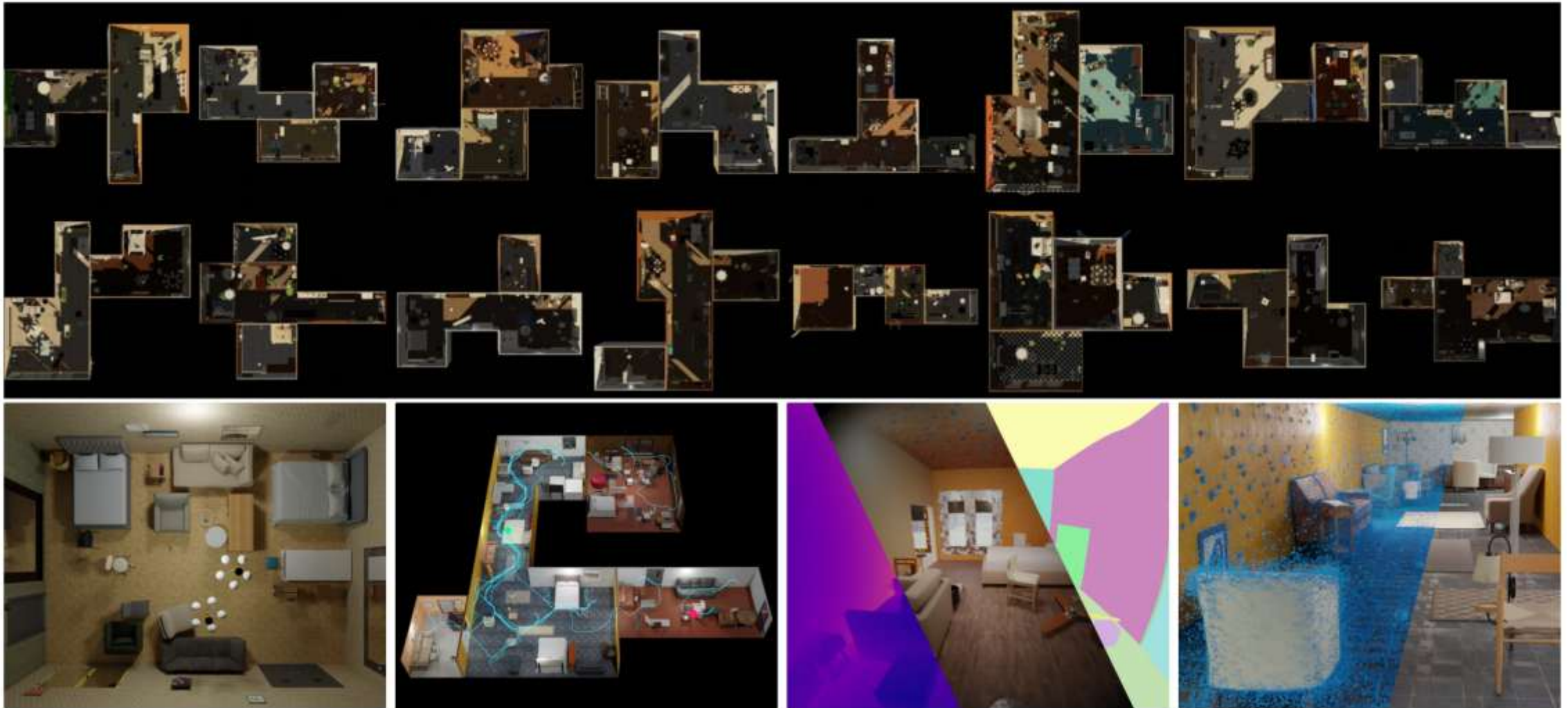
Table 1: Complete set of structured language commands designed for detailing architectural layouts and object bounding boxes. Supported data types can include `int`, `float`, `bool`. It is important to note that the language's extensibility allows for easy augmentation by introducing new commands like `make_prim`, `make_pillar`, or enhancing existing commands, such as incorporating `is_double_door` (`bool`).

make_wall (<code>int</code>)		make_door (<code>int</code>)		make_window (<code>int</code>)		make_bbox (<code>int</code>)	
id	<code>int</code>	id	<code>int</code>	id	<code>int</code>	id	<code>int</code>
a_x	<code>float</code>	wall0_id	<code>int</code>	wall0_id	<code>int</code>	class	<code>int</code>
a_y	<code>float</code>	wall1_id	<code>int</code>	wall1_id	<code>int</code>	position_x	<code>float</code>
a_z	<code>float</code>	position_x	<code>float</code>	position_x	<code>float</code>	position_y	<code>float</code>
b_x	<code>float</code>	position_y	<code>float</code>	position_y	<code>float</code>	position_z	<code>float</code>
b_y	<code>float</code>	position_z	<code>float</code>	position_z	<code>float</code>	angle_z	<code>float</code>
b_z	<code>float</code>	width	<code>float</code>	width	<code>float</code>	scale_x	<code>float</code>
height	<code>float</code>	height	<code>float</code>	height	<code>float</code>	scale_y	<code>float</code>
						scale_z	<code>float</code>

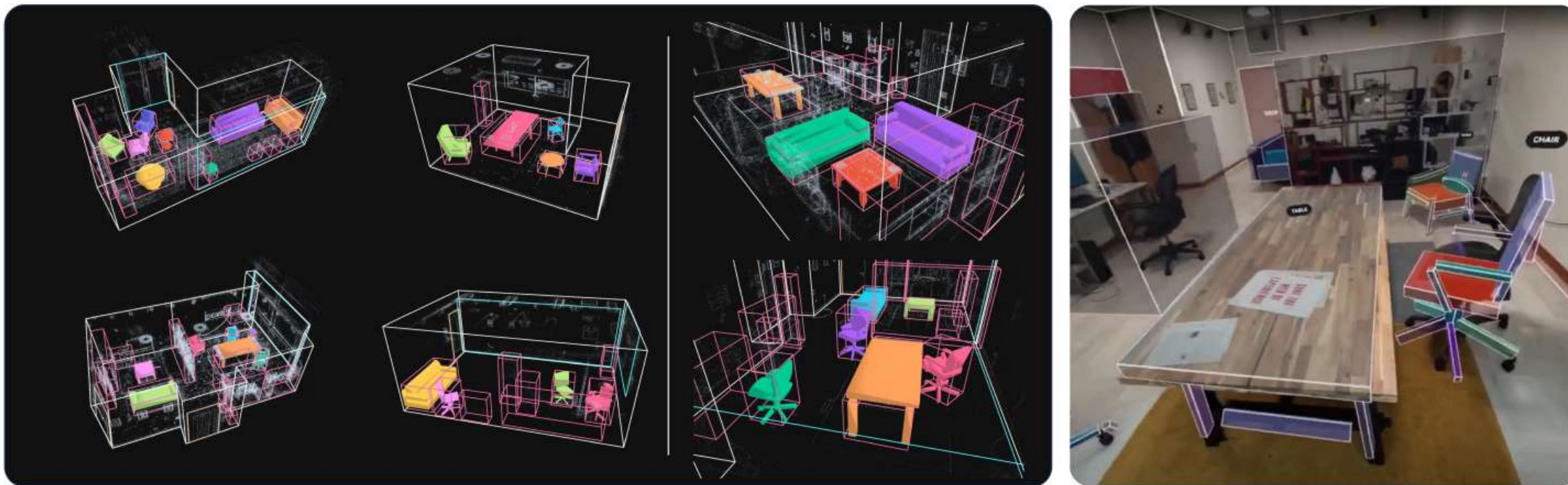
Архитектура модели



Синтетический датасет



Развитие в направлении декомпозиции объектов



```
make_prim: bbox_id, prim_num, class, center_x, center_y  
          , center_z, angle_x, angle_y, angle_z, scale_x,  
          scale_y, scale_z
```




- 3D зрение – это семантическое зрение + метрическое зрение
- Для получения 3D данных лучше всего использовать специальные сенсоры глубины и лазерные сканеры, что ограничивает доступные датасеты
- Имеющиеся 3D датасеты обогащаются текстовыми данными, что превращает их в мультимодальные и актуализирует применение visual-language моделей
- 3D LLM начали развиваться совсем недавно, опираются на обычные LLM и VLLM
- Ключевые вопросы – это подготовка датасетов и как 3D информацию закодировать в привычные токены
- Появляется возможность использовать языки программирования для конструктивного описания 3D моделей и пространства