Многомодальные системы анализа и синтеза информации в задачах управления подвижными объектами

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН, директор по направлению – руководитель научного комплекса «Искусственный интеллект и техническое зрение» ФАУ «ГосНИИАС», *viz@gosniias.ru*

XIV Всероссийское совещание по проблемам управления (ВСПУ-2024)

Москва, ИПУ РАН, 20.06.2024

Многомодальные системы анализа и синтеза информации
в задачах управления подвижными объектами

1. Концептуальный обзор современных подходов к обработке и синтезу многомодальной/многоспектральной информации

- Многомодальные данные, задачи, модели и методы ИИ
- Модели и методы анализа данных
- Модели и методы синтеза данных
- Модели и методы синтеза управления, алгоритмов и принятия решений

2. Работы и результаты ГосНИИАС в области анализа и синтеза многомодальной/многоспектральной информации

- Семантико-морфологический анализ изображений
- Анализ и синтез многоспектральных изображений для САУ ЛА
- Распределённые многоспектральные сенсоры на основе БЛА и РТК
- Интеллектуализация борта в гражданской авиации
- Анализ многомодальной информации для выработки управления
- Программно-алгоритмические средства

Концептуальный обзор современных подходов к обработке и синтезу многомодальной и многоспектральной информации

Многомодальные данные, задачи анализа и синтеза информации при создании систем управления

Модели и методы ИИ

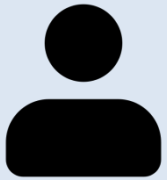
Многомодальные данные: модальности для человека и компьютера

ВикипедиЯ
Свободная энциклопедия

Мода́льность (от лат. modus — способ) — принадлежность отражаемого раздражителя к определённой сенсорной системе.

Список модальностей:

- Визуальная
- Аудиальная
- Кинестетическая
- Болевая
- другие



...Модальные характеристики ощущений отражают свойства реальности **в специфически закодированной форме**. Понятие модальности относится и ко многим другим психическим процессам, описывая **качественные характеристики когнитивных (познавательных) образов** любого уровня и сложности.

Модальность:
- Специфика сенсорной системы
- Специфика формы кодирования данных

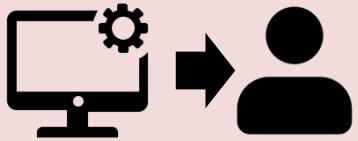
Computer–Human modalities
(send information to humans)

Common modalities

- Vision – typically through a screen
- Audition – various audio outputs
- Tactition – vibrations or movement

Uncommon modalities

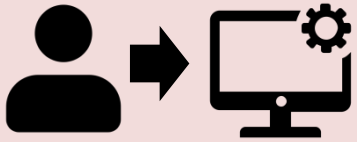
- Gustation (taste)
- Olfaction (smell)
- Thermoception (heat)



Human–computer modalities
(receive information from humans)

Simple modalities

- Keyboard
- Pointing device
- Touchscreen



Complex modalities

- **Computer vision**
- **Speech recognition**
- **Motion estimation**
- **Orientation estimation**

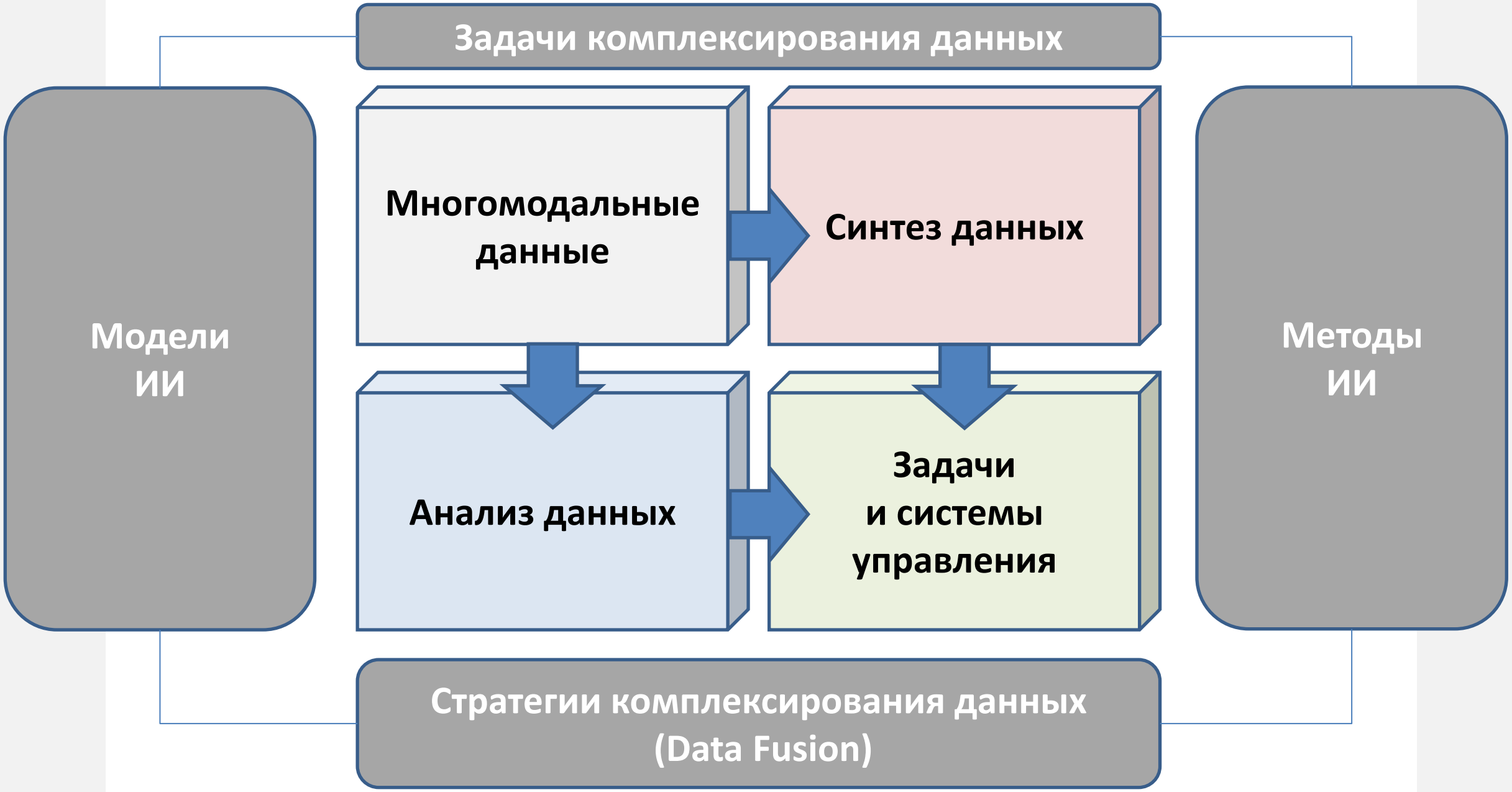
Multimodal learning, in the context of machine learning, is a type of deep learning using **multiple modalities** of data, such as **text, audio, or images**.

In contrast, unimodal models can process only one type of data, such as text or images. Multimodal learning is different from combining unimodal models trained independently. It **combines information from different modalities in order to make better predictions**.

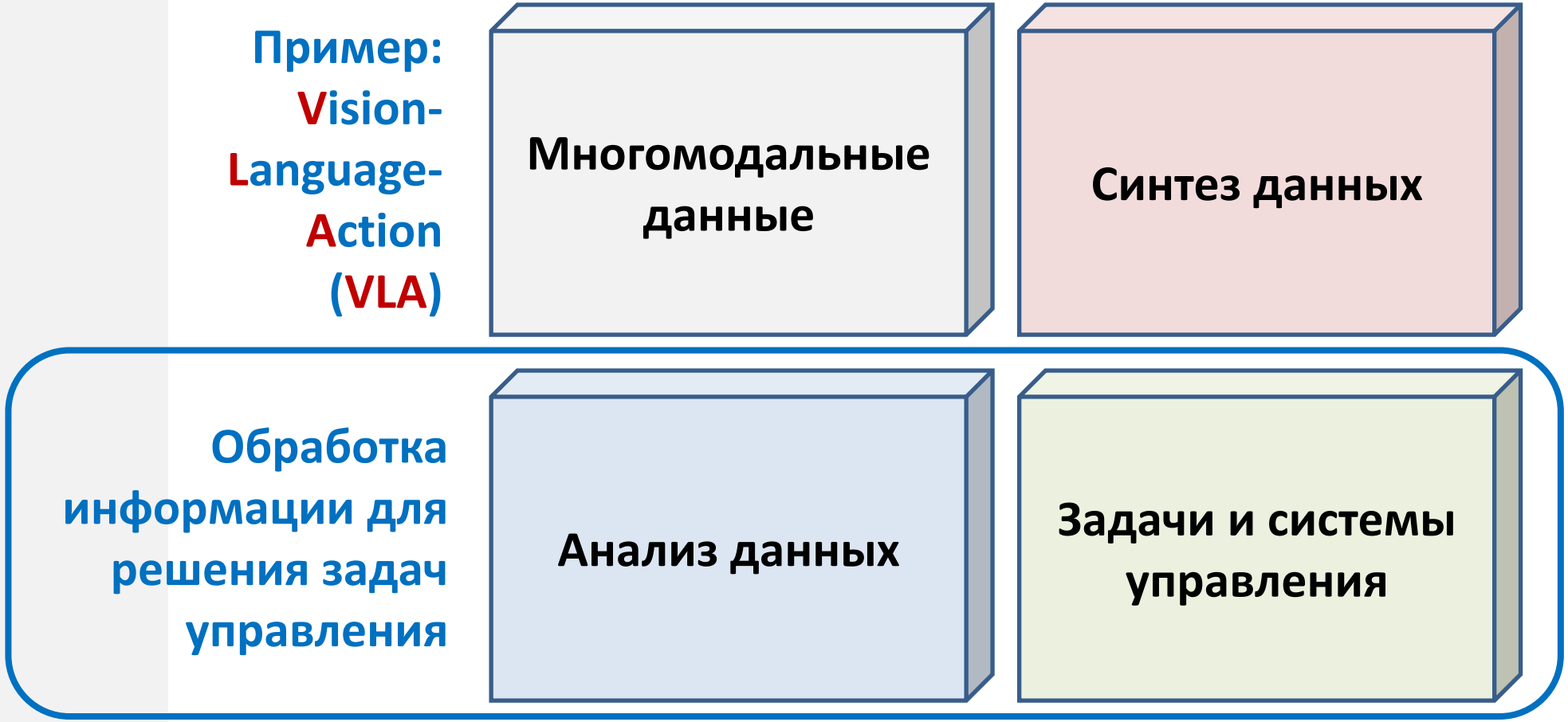
Многомодальные данные: от главных модальностей до частных доменов

Модальности	Табличные и другие цифровые данные	Сигналы, временные ряды	Пространственные данные (зрение)	Текст, символы
Субмодальности		Аудио Сигналы Речь Ряды	Размерности: 2D, 3D, 2D+T,… Структуры: Тензор, Граф, Облако точек.. Природа: ТВ, ИК, УФ, РЛ, ЛЛ…	Естественный язык Искусственный язык (программы, команды управления) Формулы и символьные выражения
Модальность: ***- Специфика сенсорной системы*** ***- Специфика формы кодирования данных***	***Проблема многомодальности технически возникает тогда, когда данные разных модальностей, структуры, природы нуждаются в специальном преобразовании для совместной обработки или генерации***			
Домены		**Разные датчики**	**Разные ракурсы**	**Разные языки**

Многомодальные данные: анализ и синтез для систем управления

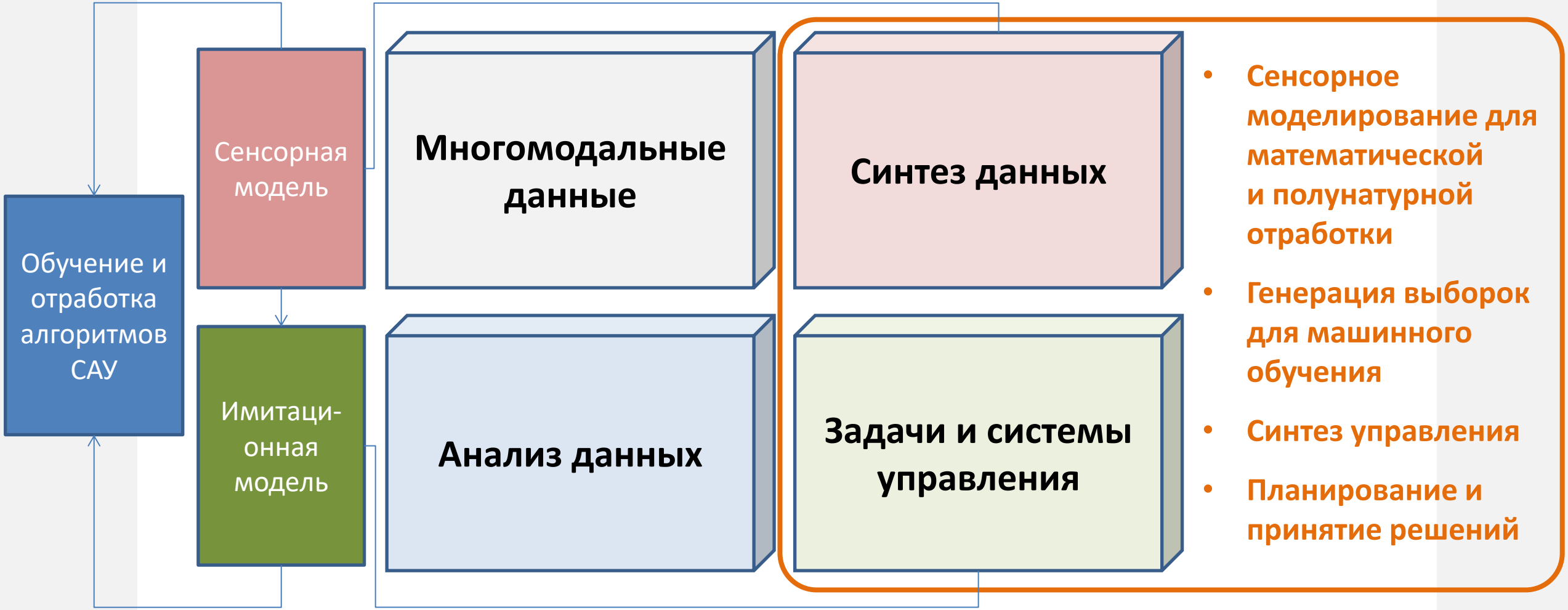


Многомодальные данные: задачи комплексирования данных



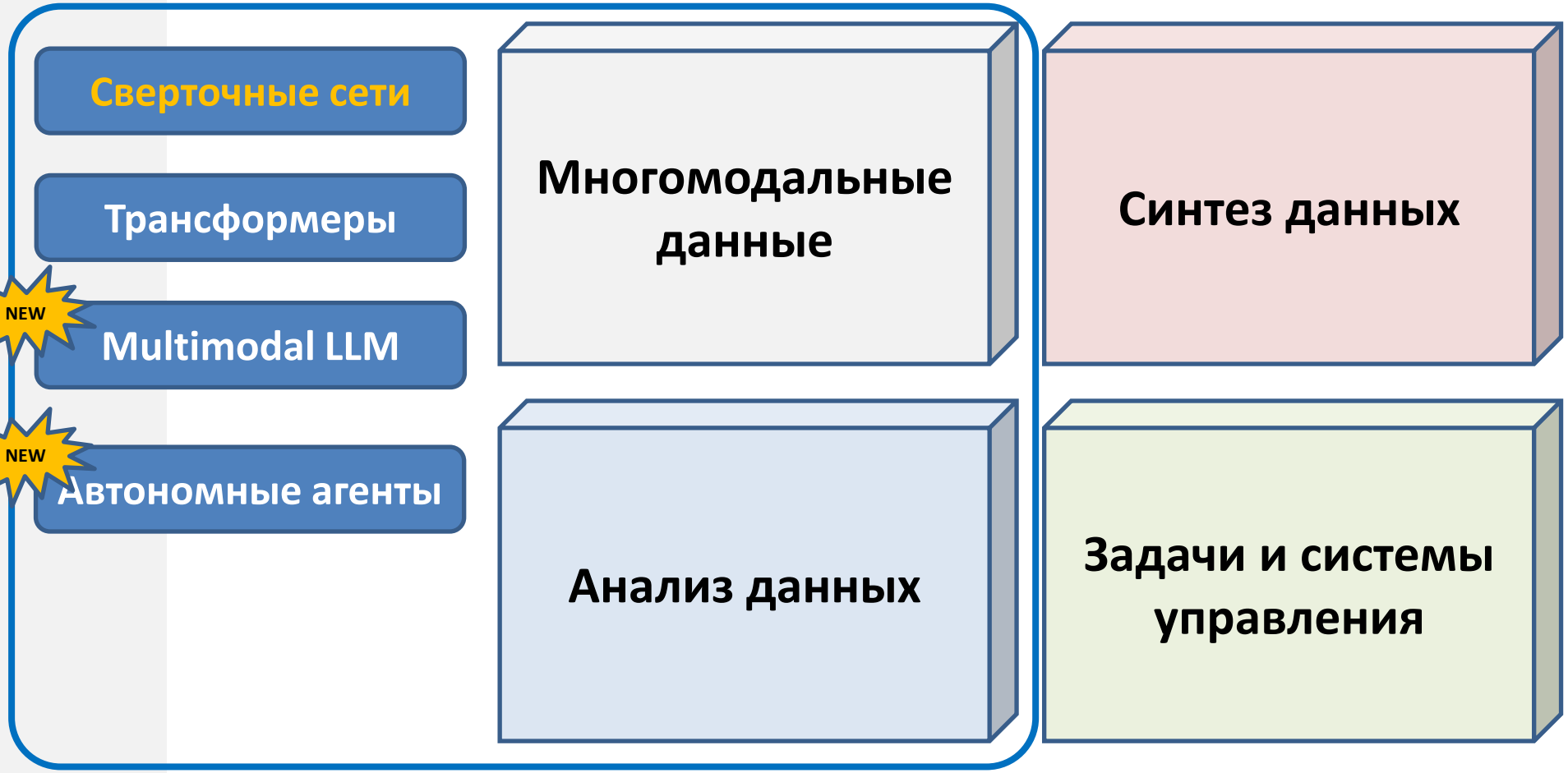
Стратегии комплексирования данных (Data Fusion):
двухэтапные: обработка информации + управление,
одноэтапные: универсальные модели для обработки информации и управления

Многомодальные данные: задачи комплексирования данных



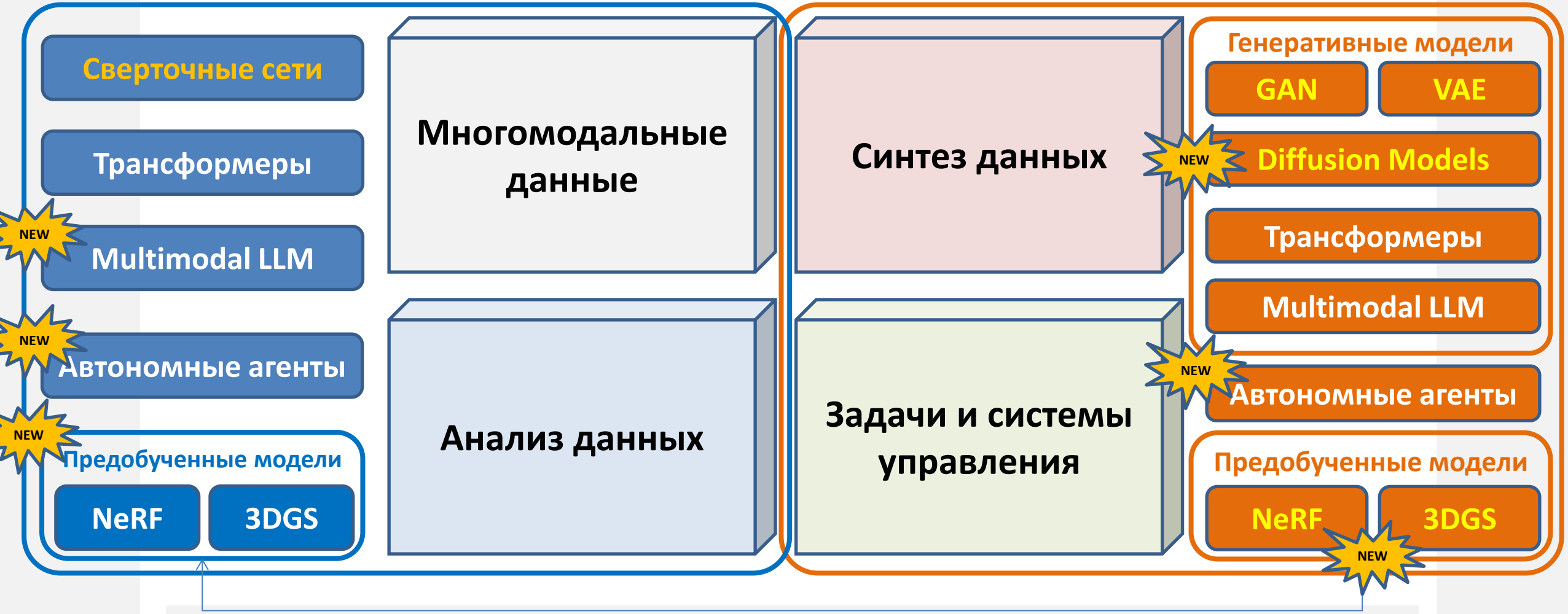
Стратегии комплексирования данных (Data Fusion):
двухэтапные: обработка информации + управление,
одноэтапные: универсальные модели для обработки информации и управления

Многомодальные данные: типы моделей для анализа и синтеза



Стратегии комплексирования данных (Data Fusion):
на уровне данных, на уровне признаков, на уровне токенов,
на уровне блоков моделей, на уровне языка, на уровне агентов

Многомодальные данные: типы моделей для анализа и синтеза



Стратегии комплексирования данных (Data Fusion):
на уровне данных, на уровне признаков, на уровне токенов,
на уровне блоков моделей, на уровне языка, на уровне агентов

Многомодальные данные: методы ИИ для анализа и синтеза

Машинное обучение

Обучение с учителем (Supervised)

Обучение без учителя (Self-supervised)

Инженерия запросов (Prompt Engineering)

NEW

Использование обученных моделей для ответа на многомодальные запросы

Многомодальные данные

Синтез данных

Анализ данных

Задачи и системы управления

Стратегии комплексирования данных (Data Fusion):
до обучения, в процессе обучения,
в процессе дообучения, после обучения

Многомодальные данные: методы ИИ для анализа и синтеза

Машинное обучение

- Обучение с учителем (Supervised)
- Обучение без учителя (Self-supervised)
- Инженерия запросов (Prompt Engineering) — NEW

Использование обученных моделей для ответа на многомодальные запросы

Многомодальные данные

Анализ данных

Синтез данных

Задачи и системы управления

Машинное обучение

- Обучение с учителем (Supervised)
- Обучение без учителя (Self-supervised)
- Обучение с подкреплением (Reinforcement Learning)
- Инженерия запросов (Prompt Engineering) — NEW

Использование обученных моделей для генерации/принятия решений/управления

Стратегии комплексирования данных (Data Fusion):
до обучения, в процессе обучения,
в процессе дообучения, после обучения

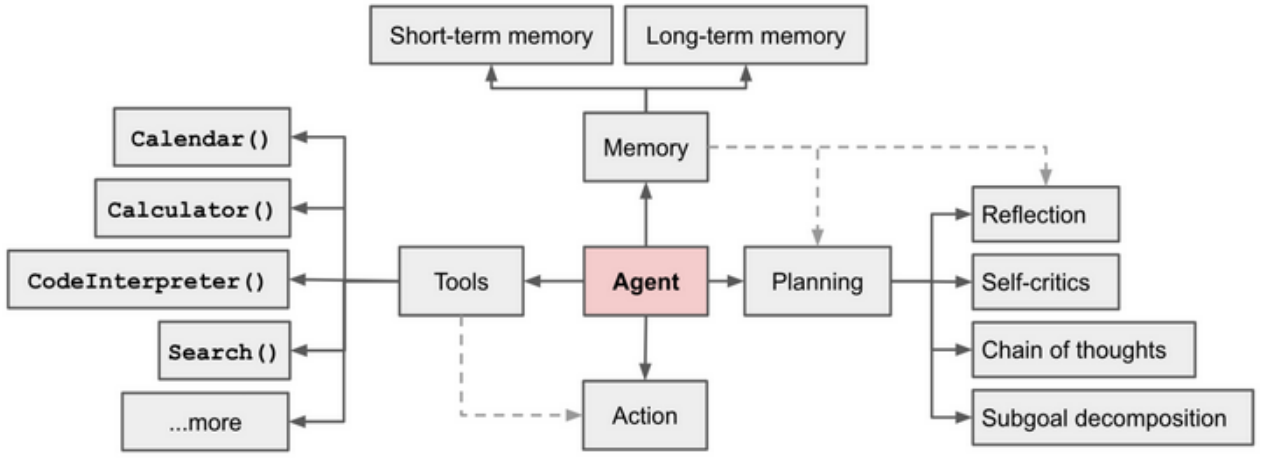
Large Language Models (LLM), LLM Powered Autonomous Agents (LLM-агенты)

Современный ландшафт LLM

2021	2022	2023	2023
		Claude	Gemini
Jurassic-1	CodeGen	Vicuna	CodeGen 2
Retro	BLOOM	Mistral	StarCoder
FLAN	ChinChilla	Alpaca	Grok
Web-GPT	Instruct GPT	GPT-4	PaLM-2
T0	OPT	DPO	Toolformer
Ernie 3.0	Galactia	Llama 1/2	Zephyr
CODEX	PaLM	Phi-1/2	Mixtral
Gopher	LaMDA	FALCON	Mamba-Chat
		MPT	ORCA-2

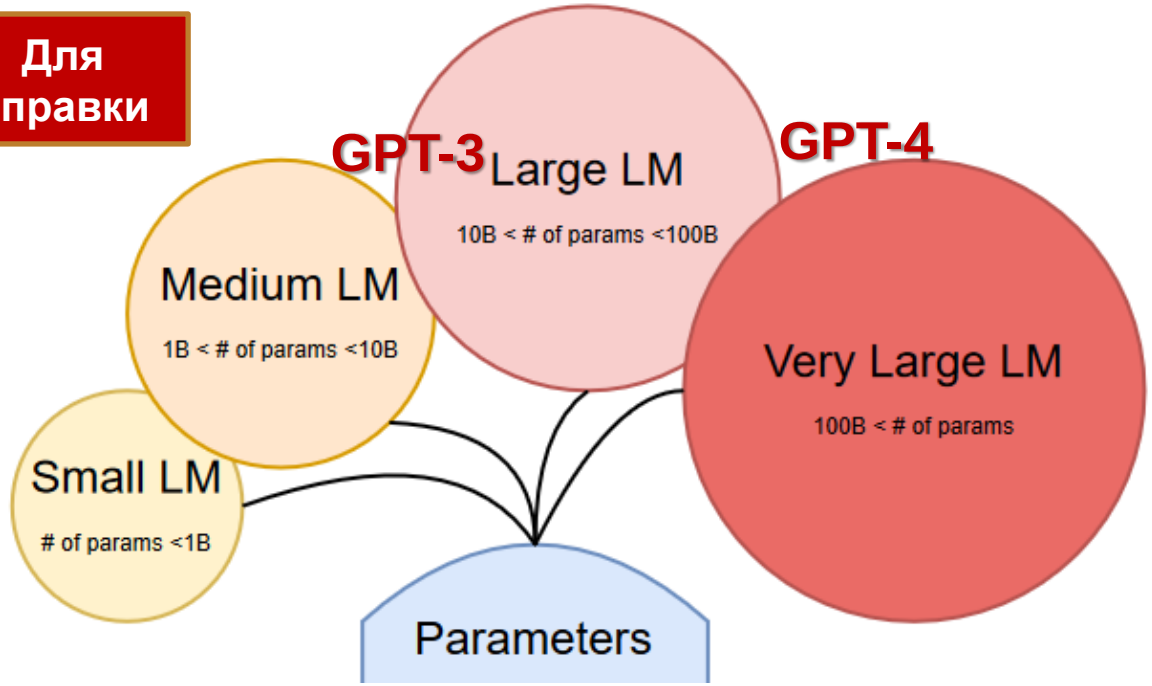
*Термины, связанные с LLM

Модели разных размеров



LLM Powered Autonomous Agents

Для справки



Planning, subgoal and decomposition: The agent breaks down large tasks into smaller, manageable subgoals, enabling efficient handling of complex tasks.

Reasoning, reflection and refinement: The agent can do self-criticism and self-reflection over past actions, learn from mistakes and refine for future steps, thereby improving the quality of final results.

Tool use: The agent learns to call external APIs for extra information that is missing from the model weights, including current information, code execution capability, access to proprietary information sources and more.

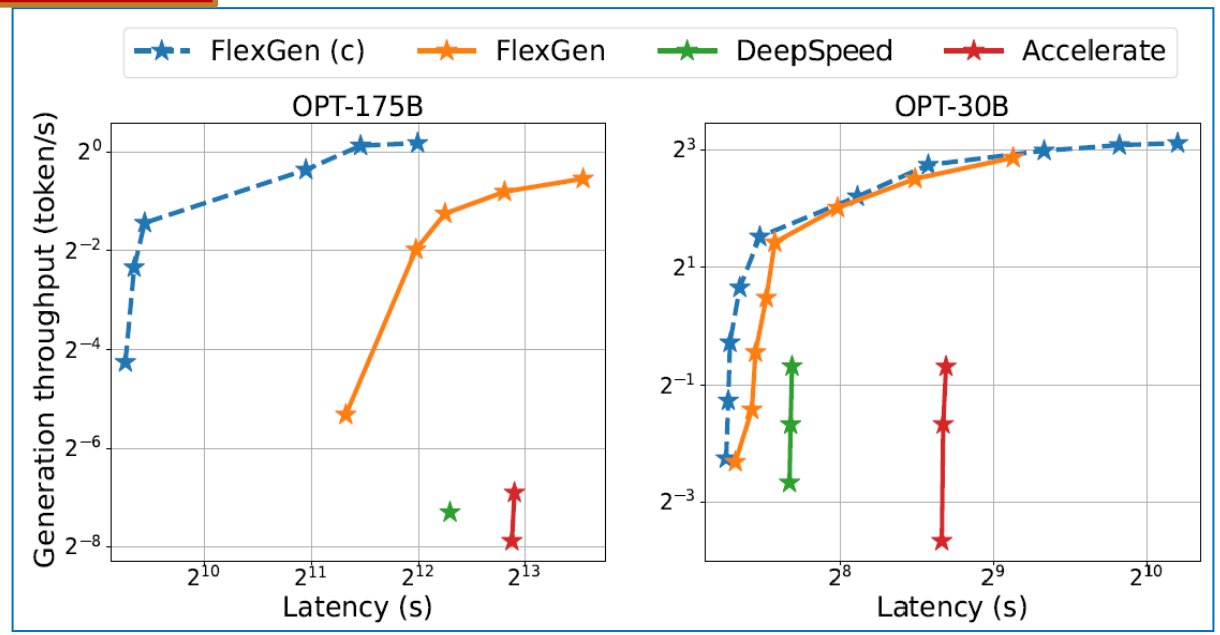
С точки зрения LLM, агенты – всего лишь способ их промптинга. С точки зрения агентов – LLM всего лишь один инструмент для решения задач

Large Language Models: A Survey, 2024

Weng, Lilian. (Jun 2023). “LLM-powered Autonomous Agents”. Lil’Log.

Для справки

Можно ли использовать LLM в бортовых САУ? ДА!

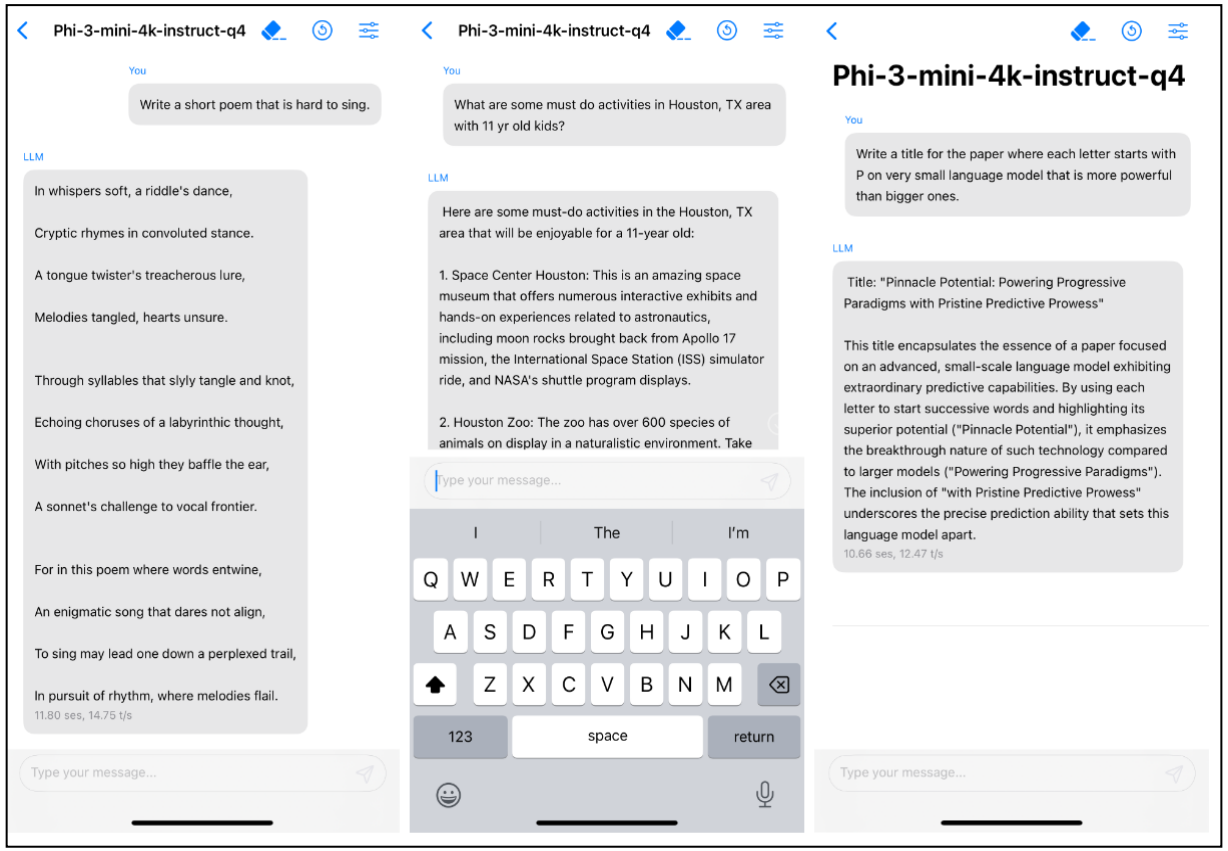


175B параметров = GPT-3

Running **OPT-175B on a single 16GB GPU**, FlexGen achieves **a generation throughput of 1 token/s for the first time with an effective batch size of 144**.

Обучение LLM с нуля простым смертным пока недоступно. Дообучение возможно! Использование вполне возможно!

High-throughput Generative Inference of Large Language Models with a Single GPU, март **2023**



4-bit quantized phi-3-mini running natively on an iPhone with A16 Bionic chip, generating over 12 tokens per second.

Сегодня LLM можно запустить локально даже на iPhone!

Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, апрель **2024**

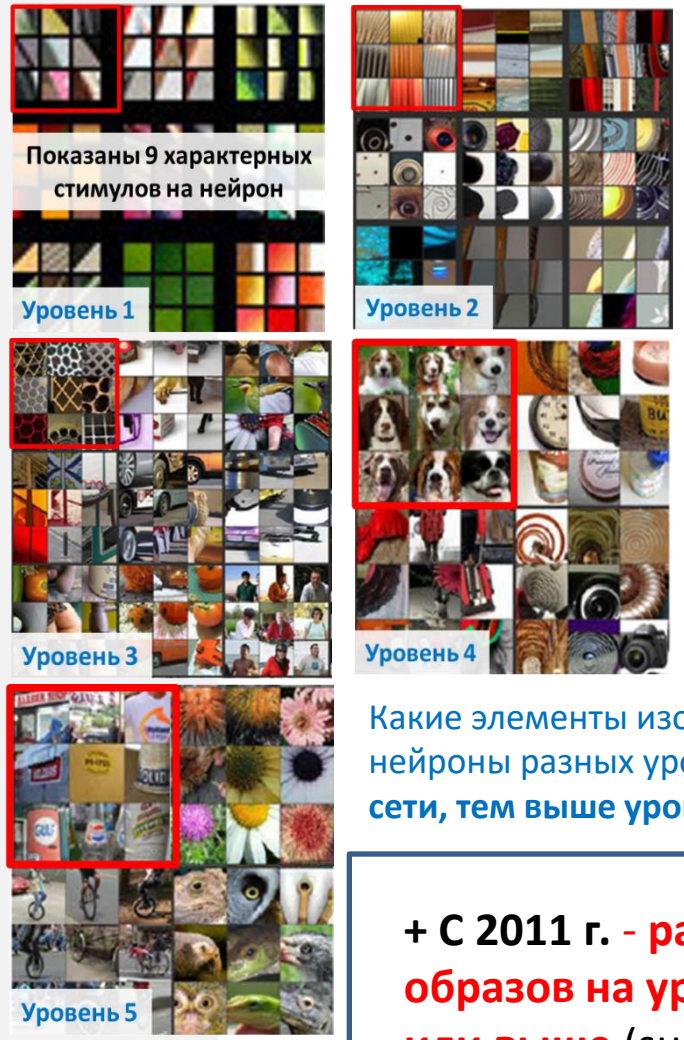
Модели и методы многомодального анализа данных

(на примере Vision-Language задач)

Исходный инструмент:
Персептрон (MLP)

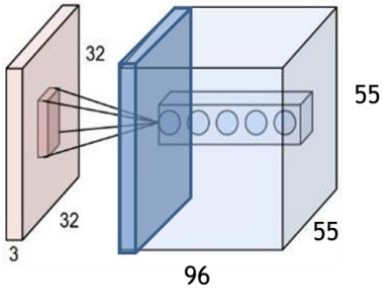
Сверточные нейронные сети для задач зрения

CV до 2020



Visualizing and Understanding Convolutional Networks,, ECCV'14

Ключевое изобретение: сверточный нейрон
Работает локально в небольшой окрестности



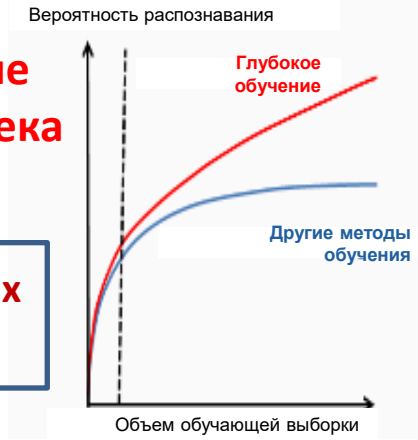
+ Иерархическая обработка с повышением абстракции данных от уровня к уровню

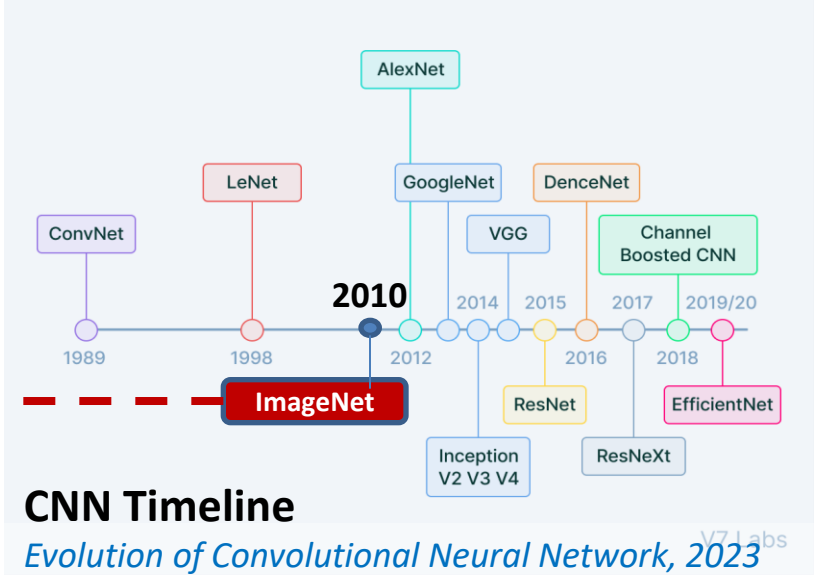
Какие элементы изображения распознают нейроны разных уровней: **чем выше слой сети, тем выше уровень абстракции**

AlexNet (2011)

+ С 2011 г. - распознавание образов на уровне человека или выше (superhuman)

+ Обучение на сверхбольших объемах данных





CNN Timeline

Evolution of Convolutional Neural Network, 2023

Object Detection with CNN Timeline

CNN Architectures

AlexNet
VGG16
VGG19
Inception V1
Inception V2
GoogleNet
ResNet18
ResNet34
ResNet50
ResNet101
ResNet152
DenseNet
SqueezeNet
MobileNet V2
ShuffleNet V2
EfficientNet

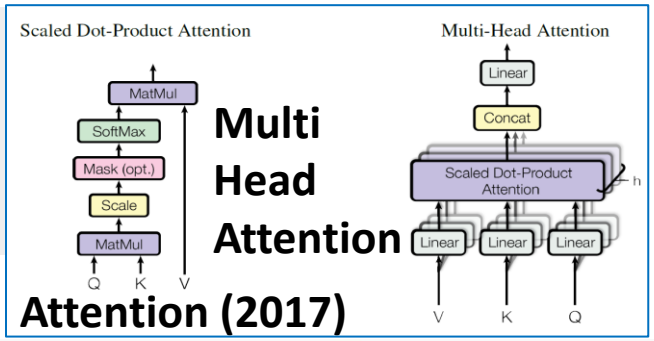
Основное ограничение: нет символьных представлений, нет структурных моделей, нет логик и онтологий

Важная проблема: необходимость обучения на больших выборках (**сложно Zero-Shot/ Few-Shot**)

Трансформеры для анализа естественного языка

NLP до 2020

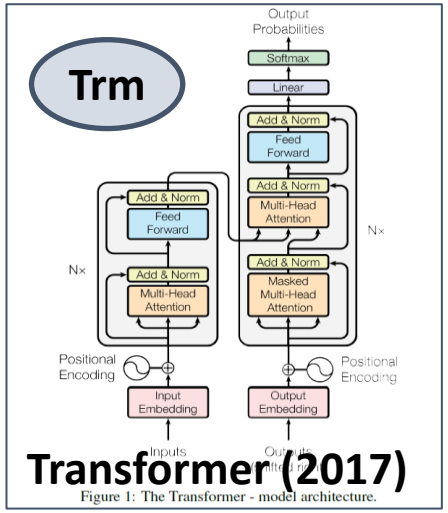
Attention Is All You Need (Vaswani et.al., 2017)



Multi Head Attention

Attention (2017)

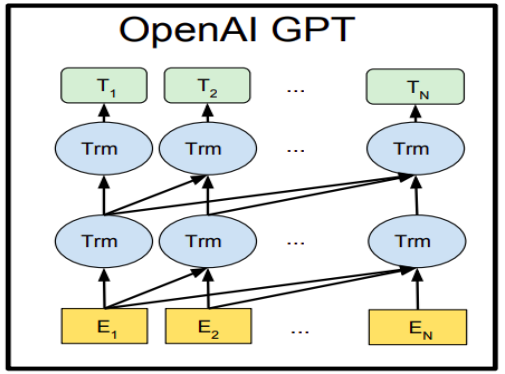
Модуль внимания работает нелокально, выделяя смысловые связи между словами, далеко отстоящими в тексте



Transformer (2017)

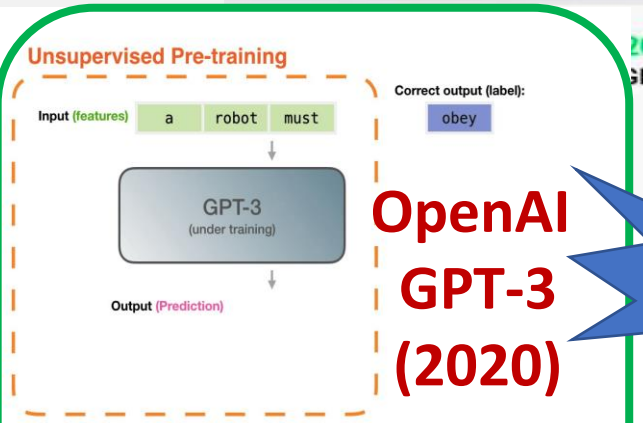
Figure 1: The Transformer - model architecture.

Generative Pre-trained Transformer (GPT, 2018)



Self-supervised переобучение (без разметки данных!)

Нейронные сети в NLP

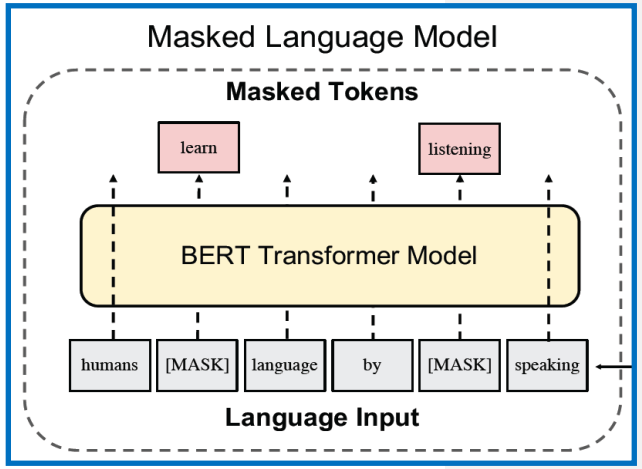


OpenAI GPT-3 (2020)

ХИТ ИИ 2020!

Новый класс моделей: Большие языковые модели (LLM)

GPT учится предсказывать следующий токен текста



The CommonCrawl data - 45 Тб
SOTA в 42 задачах NLP

GPT-3 может отвечать на вопросы, переводить с разных языков, писать статьи, исходные коды программ…

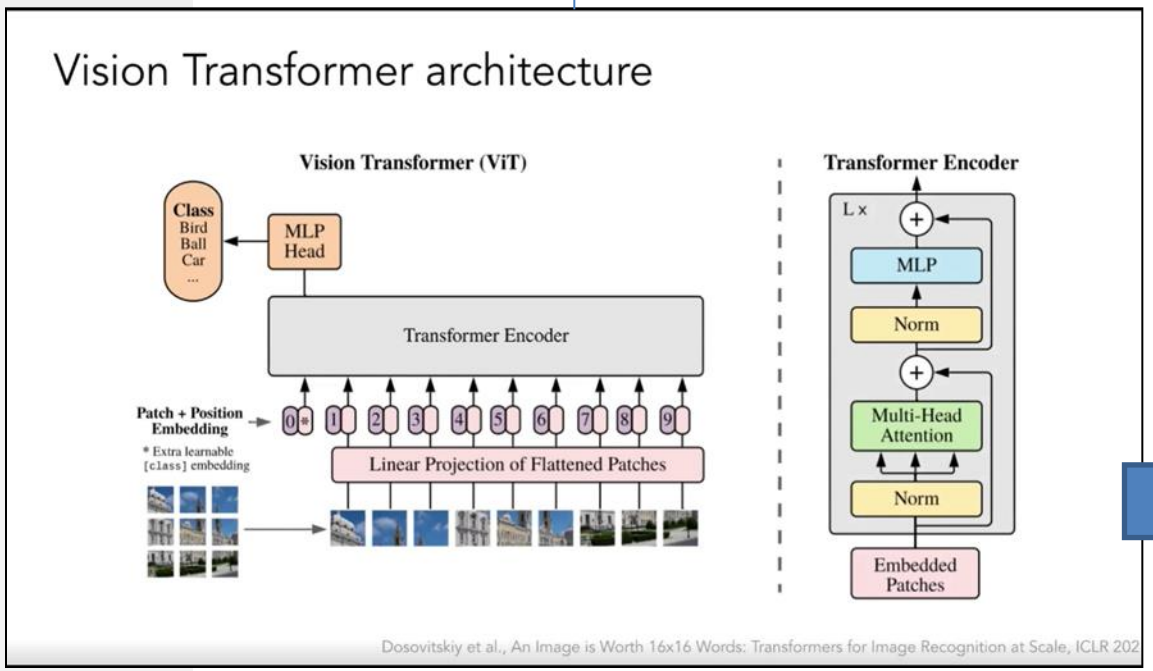
Основное ограничение: только символьные данные.
Не работают с изображениями и данными других модальностей

Трансформеры идут в CV!

CV+NLP 2020+

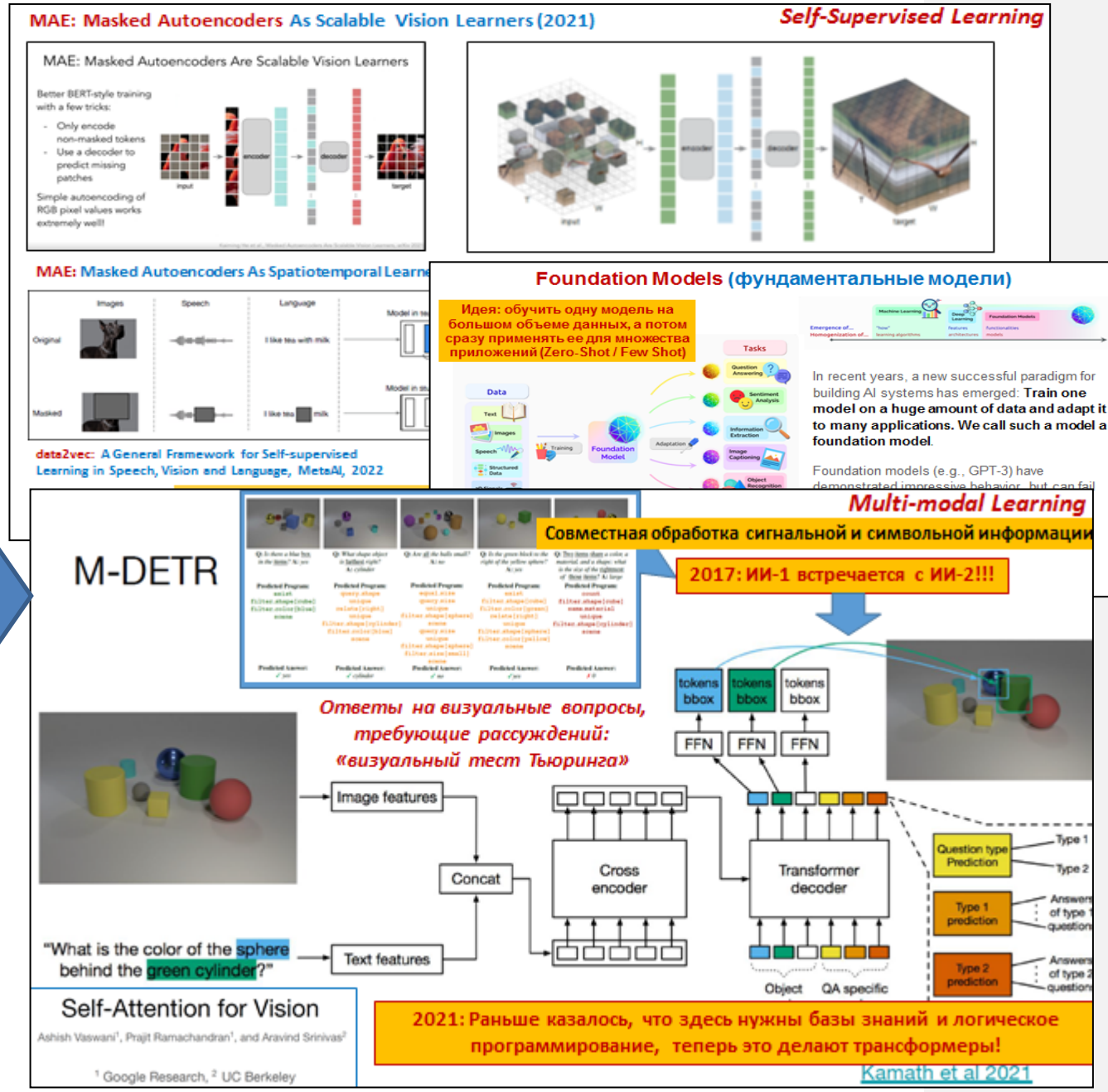
Базовая идея применения трансформеров к задачам зрения: вместо текстовых токенов - фрагменты изображений

Vision Transformer (2020)



Изображение как «текст»: все токены сравниваются попарно, отсюда квадратичные вычисления

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale ***Alexey Dosovitskiy***, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby

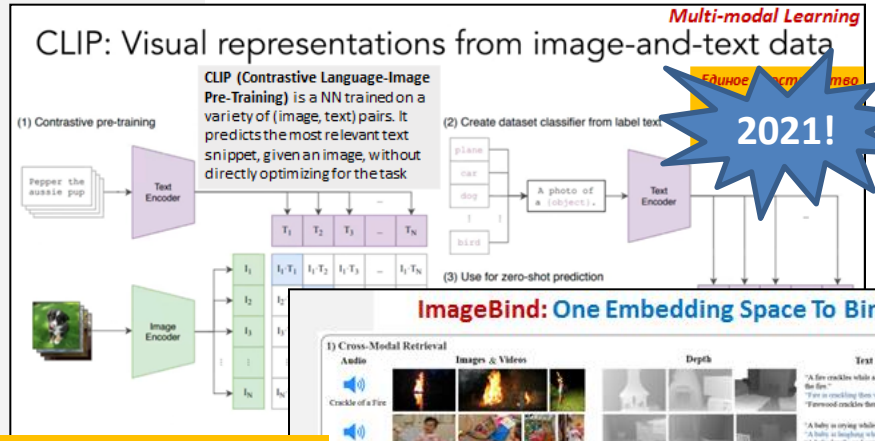


Multimodal Learning
Vision-Language Models (VLM)

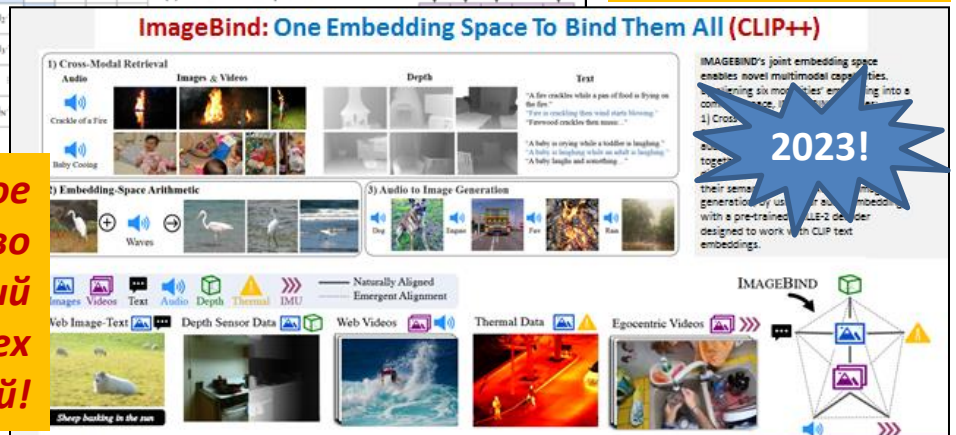
Multimodal Learning

Overvew

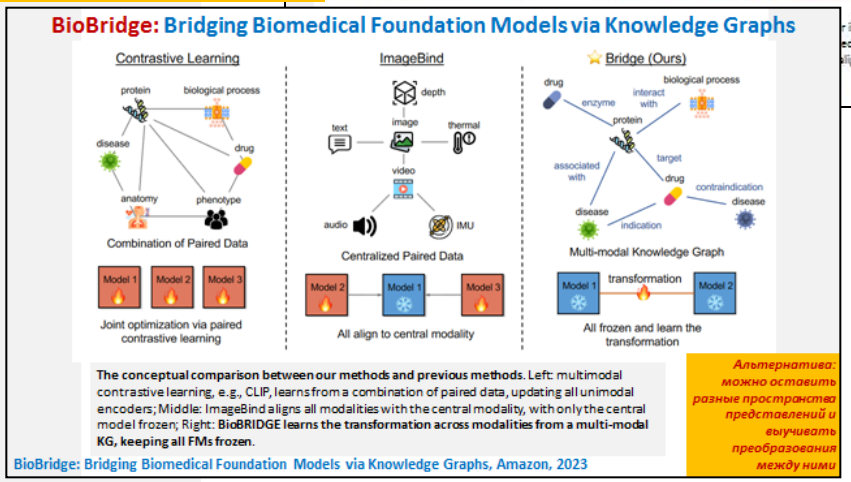
Data Fusion at the feature level (на уровень ниже)



Единое пространство представлений для текста и изображений!



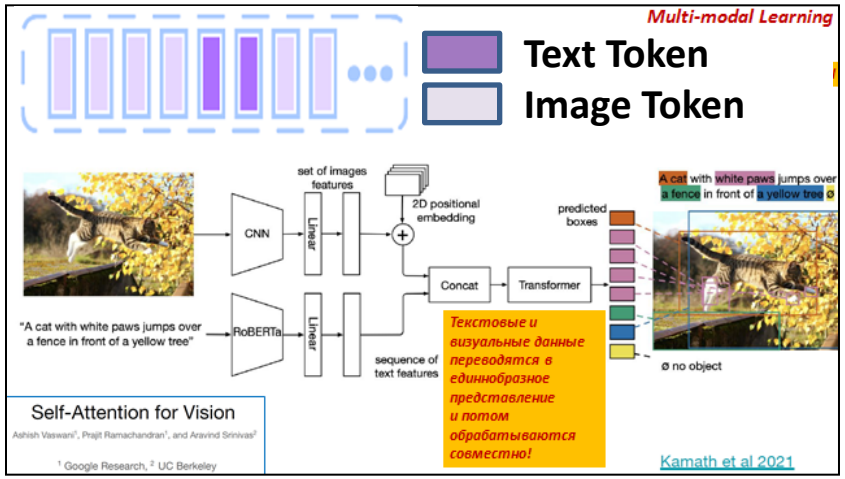
Единое пространство представлений для всех модальностей!



Альтернатива: можно оставить разные пространства представлений и выучивать преобразования между ними

Разные пространства представлений и выученные преобразования между ними

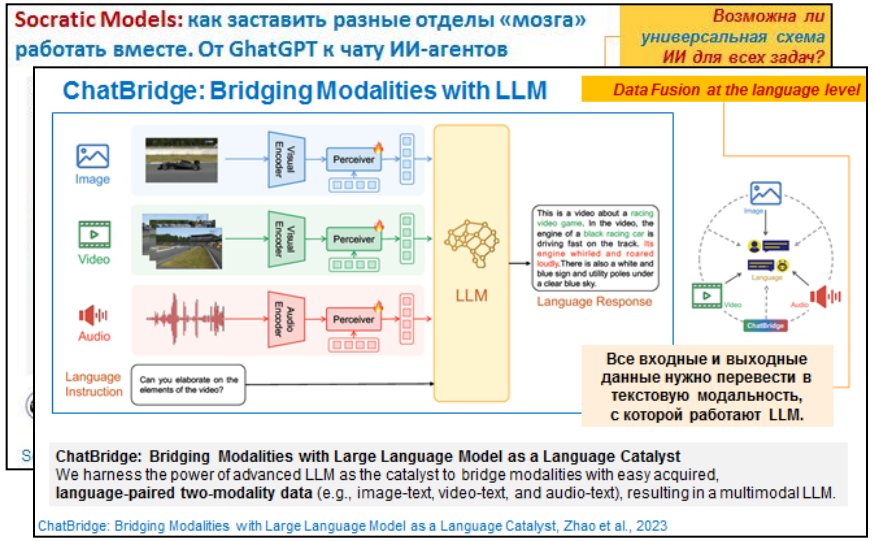
Data Fusion at the token level



Текстовые и визуальные данные переводятся в единнобразное представление и потом обрабатываются совместно!

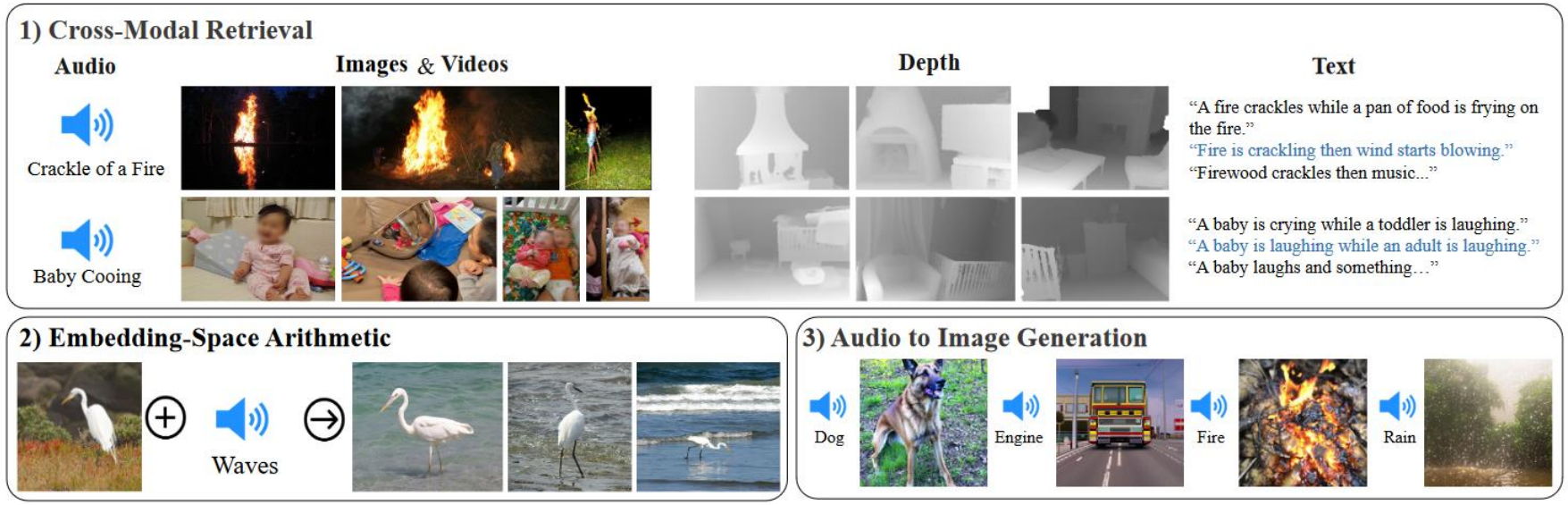
Текстовые и визуальные данные переводятся в токены, единая лента которых идет на вход трансформера

Data Fusion at the language level (на уровень выше)

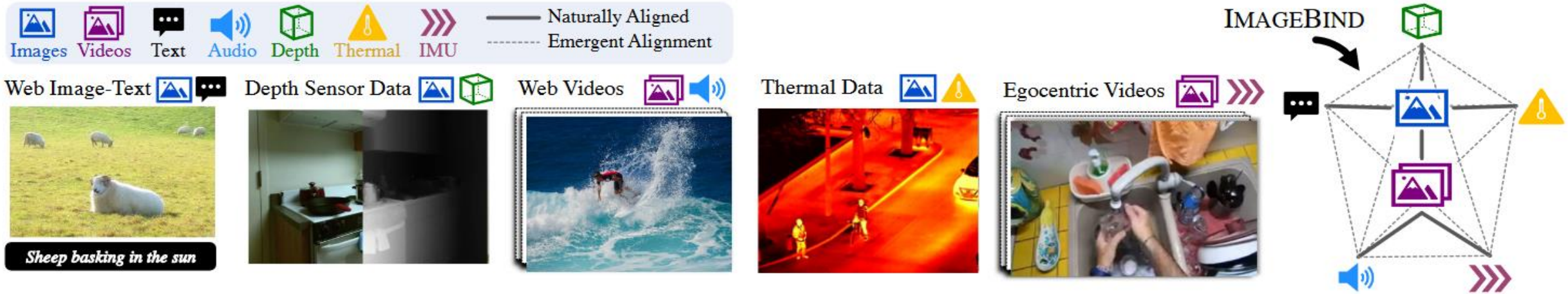


Все входные и выходные переводим в текстовую модаль-ность, с которой работают LLM в чате

ImageBind: One Embedding Space To Bind Them All (CLIP++)



IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: 1) Cross-Modal Retrieval, which shows emergent alignment of modalities such as audio, depth or text, that aren't observed together. 2) Adding embeddings from different modalities naturally composes their semantics. And 3) Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 decoder designed to work with CLIP text embeddings.



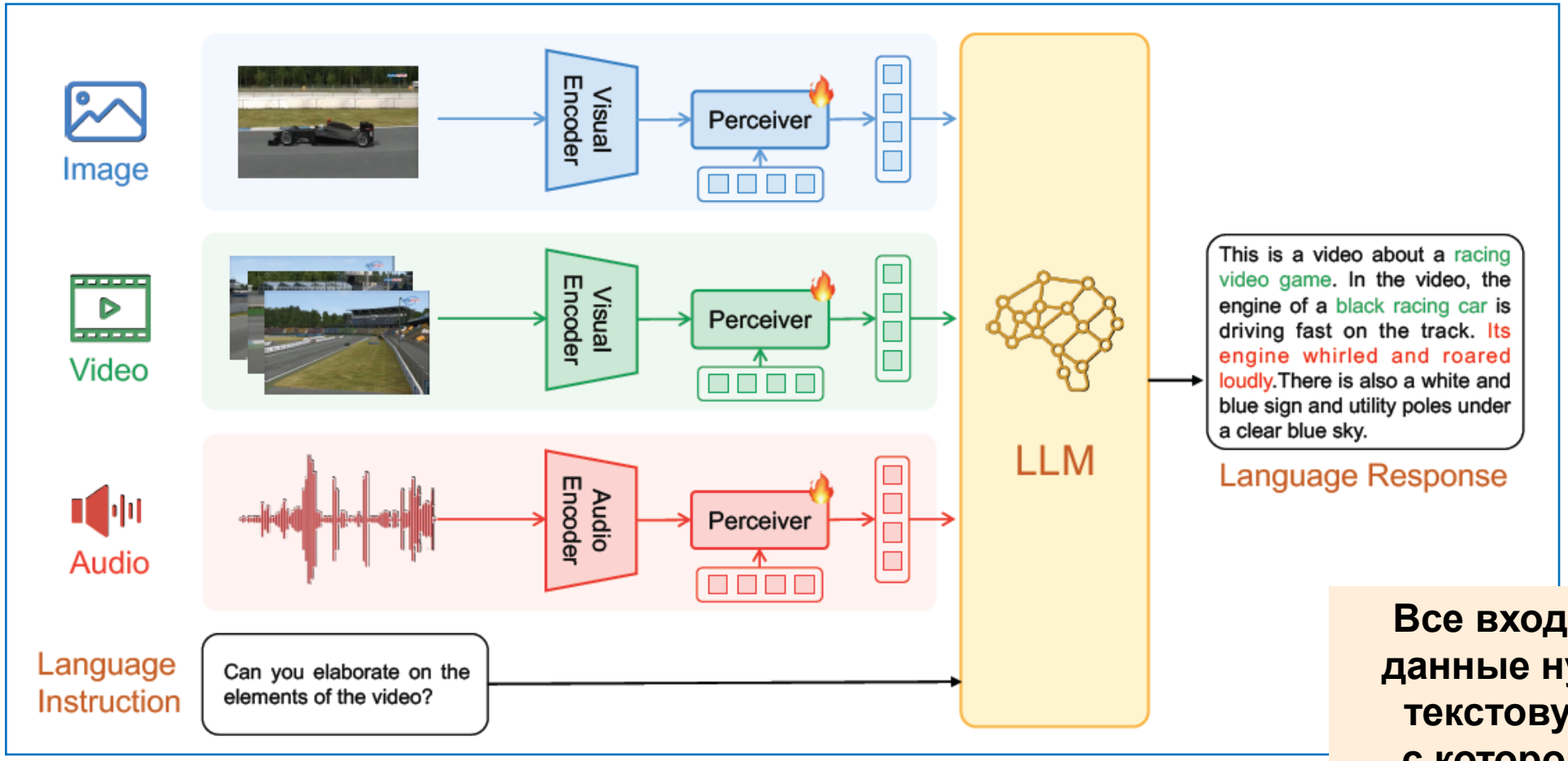
IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, etc. IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities

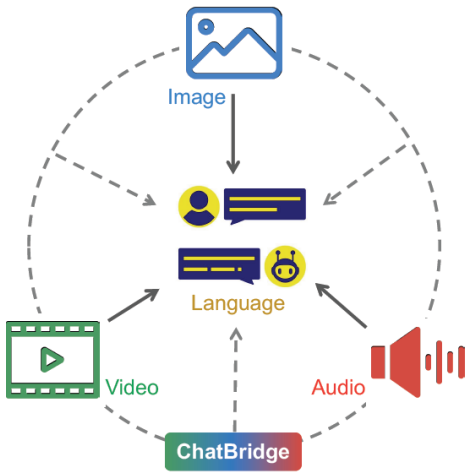
Единое пространство представлений для всех модальностей!

IMAGEBIND: One Embedding Space To Bind Them All, FAIR, Meta AI, 2023

ChatBridge: Bridging Modalities with LLM

Data Fusion at the language level



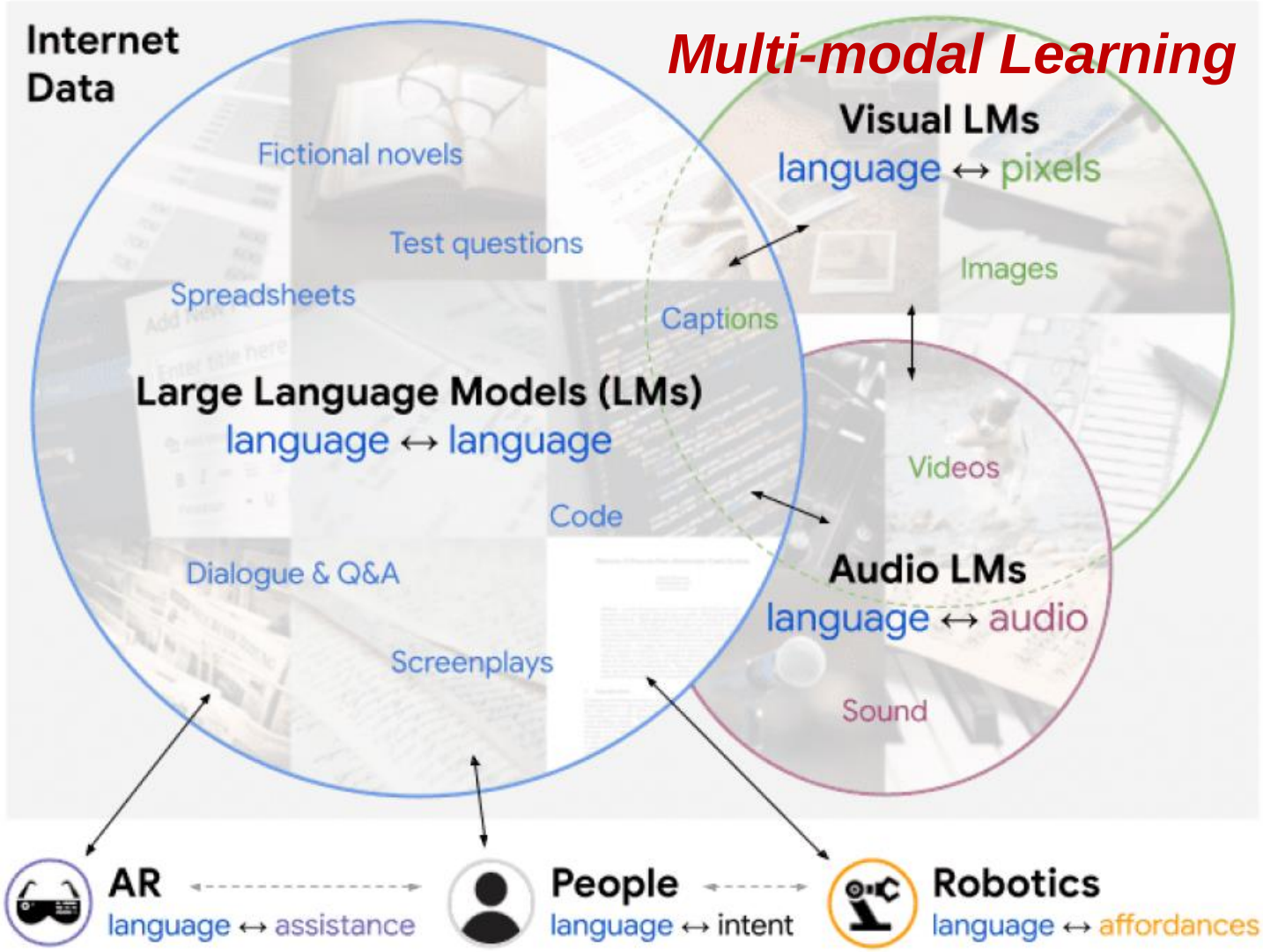


Все входные и выходные данные нужно перевести в текстовую модальность, с которой работают LLM.

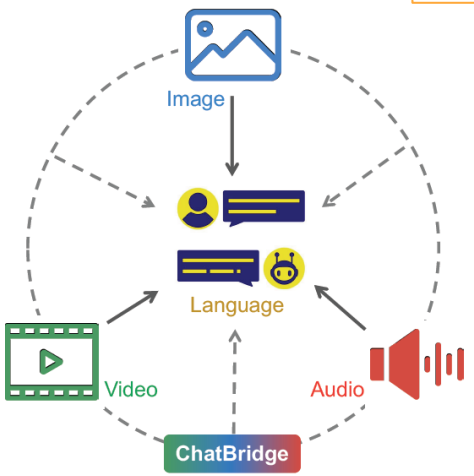
ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst
We harness the power of advanced LLM as the catalyst to bridge modalities with easy acquired, **language-paired two-modality data** (e.g., image-text, video-text, and audio-text), resulting in a multimodal LLM.

ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst, Zhao et al., 2023

Socratic Models: решение мультимодальных задач коллективом агентов. От GhatGPT к чату ИИ-агентов



Data Fusion at the agent level



Гипотеза: для этого все входные и выходные данные нужно перевести в текстовую модальность, с которой работают LLM.

Проблема: для обучения нужна огромная мультимодальная база

Решение: научить модели использовать запросы, и пусть модели общаются между собой

«Сократические модели» это способ построения расширяемого коллектива ИИ-агентов с «открытым» списком задач и типов данных, для работы с которыми добавляются новые агенты

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Google, 2022

Фундаментальные модели для CV

(Zero-Shot Solving Classical CV Tasks)

Уровни анализа данных в CV

Уровень сцены	**классификация и описание сцен (отношения, действия), VQA**
Уровень объектов	**обнаружение объектов, семантическая сегментация…**
Уровень признаков	**точки, линии, углы, области, контуры, текстура… matching, 3DRec**
Уровень изображения	**пикселы (положение, яркость, цвет)**

Fundamental Models for Vision Tasks

Overview

Supervised ***Self-supervised***

На какой задаче предобучить аундамен-тальную модель для CV?

Recognize Anything: Image Tagging (Multi-label Image Recognition)

Возможно, любая сложная визуальная задача может стать источником для

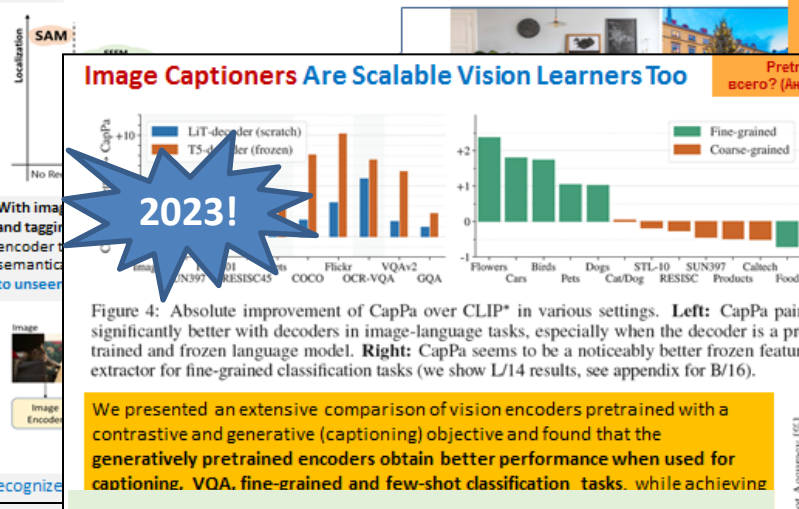


Image Captioners Are Scalable Vision Learners Too

Pretraining: Какая прокси-задача лучше всего? (Аналог классификации на ImageNet)

2023!

Figure 4: Absolute improvement of CapPa over CLIP* in various settings. **Left:** CapPa pairs significantly better with decoders in image-language tasks, especially when the decoder is a pre-trained and frozen language model. **Right:** CapPa seems to be a noticeably better frozen feature extractor for fine-grained classification tasks (we show L/14 results, see appendix for B/16).

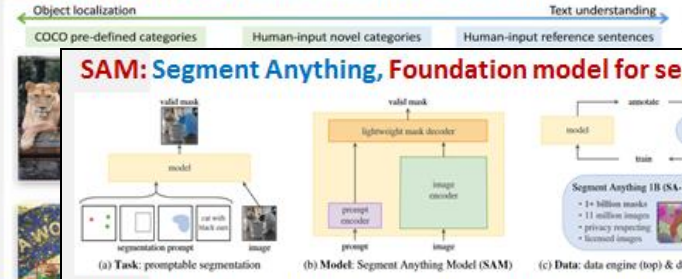
Figure 5: Absolute improvement (single-task) of CapPa over CLIP* for a decoder trained from scratch (LiT-d.), a frozen T5 decoder, and a frozen GPT-2 similar to [44].

We presented an extensive comparison of vision encoders pretrained with a contrastive and generative (captioning) objective and found that the **generatively pretrained encoders obtain better performance when used for captioning, VQA, fine-grained and few-shot classification tasks**, while achieving

Уровень сцены

Grounding DINO: Open-Set Object Detection

Задача обнаружения столь же фундаментальна!

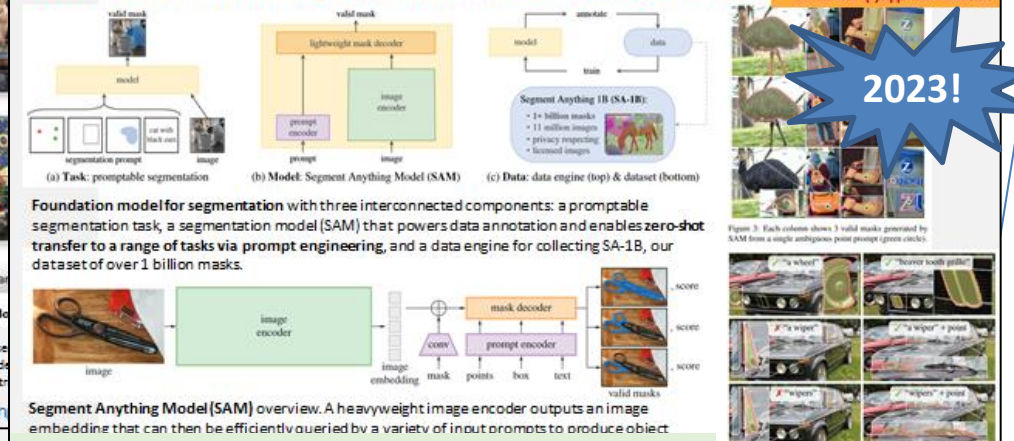


Collaborate with stable diffusion.

SAM: Segment Anything, Foundation model for segmentation

Задача сегментации столь же фундаментальна!

2023!

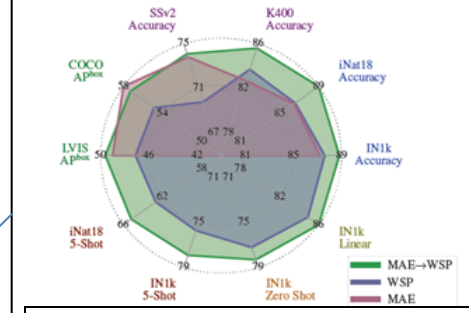


Foundation model for segmentation with three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables **zero-shot transfer to a range of tasks via prompt engineering**, and a data engine for collecting SA-1B, our dataset of over 1 billion masks.

Уровень объектов

MAE *pre*-pretraining for billion-scale pretraining

MAE + Weakly-supervised pretraining (WSP)



Pre-pretraining Masked Autoencoder (MAE) learns visual representations from image datasets without using any

MAE + Weakly-supervised pretraining (WSP)

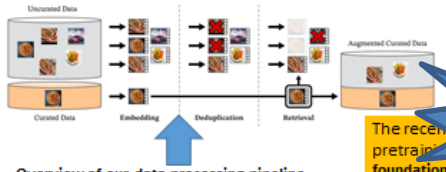
MAE→WSP, or MAWS for short, first trains the encoder using the MAE self-supervised method using only the images. This pre-pretraining stage initializes the model while simultaneously being computationally efficient

DINOv2: Learning Robust Visual Features without Supervision

Discriminative Self-supervised Pre-training

DINO: DIstillation with NO labels. DINO simplifies self-supervised ViT training by directly predicting the output of a teacher network (built with a momentum encoder) by using a standard cross-entropy loss.

Emerging Properties in Self-Supervised Vision Transformers, 2021



Overview of our data processing pipeline. Images from curated and uncurated data sources are first mapped to embeddings. Uncurated images are then deduplicated before being matched to curated images. The resulting

Discriminative Self-supervised Pre-training

2024!

The recen… language processing for model pretrai… data have opened the way for similar **foundation mo… …on.** These models could greatly simplify the use of images in any system by producing general-purpose visual features, i.e., **features that work across image distributions and tasks without finetuning.** This work shows that existing pretraining methods, especially self-supervised methods, can produce such features if trained on enough curated data from diverse sources. We revisit …bine different techniques to scale our …nd model size.

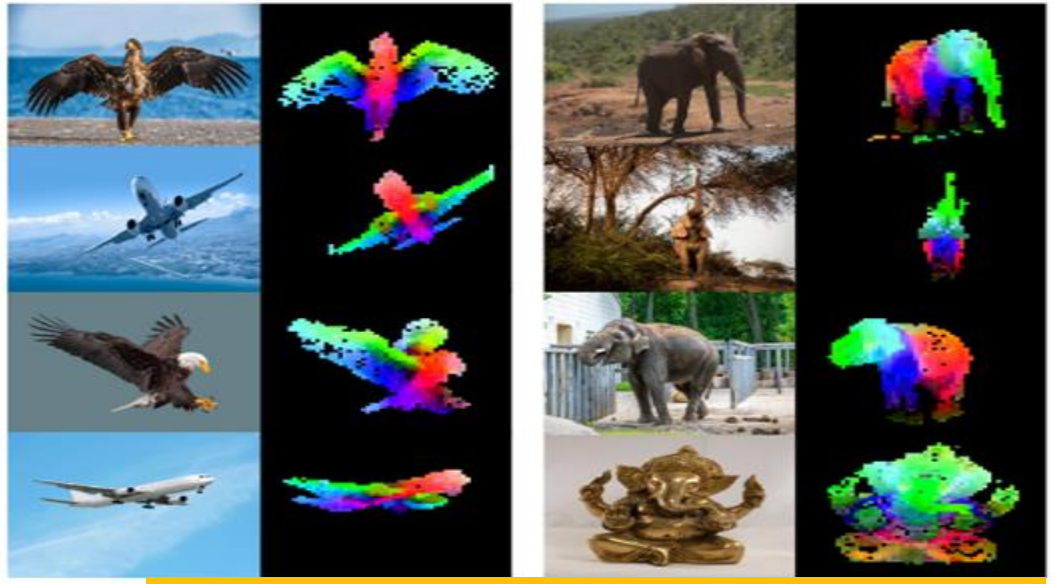
…esearch, 2024.

Уровень изображения

Видимо, любая сложная визуальная задача может стать источником для получения фундаментальной модели, которую не придется потом дообучать для решения новых задач: важно, чтобы в ней выработались нужные признаки

DINOv2: Learning Robust Visual Features without Supervision

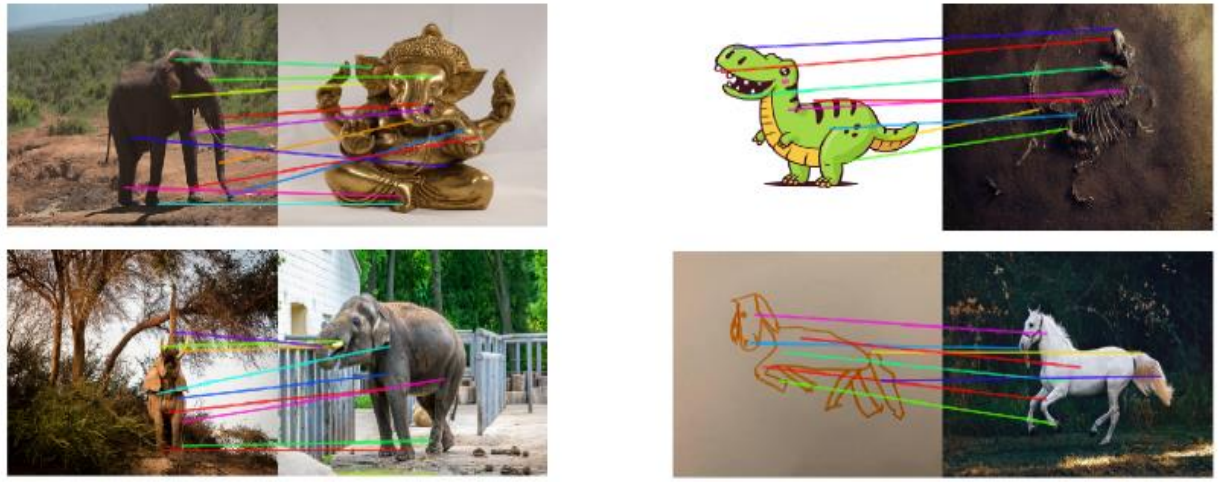
Фундаментальная модель должна иметь универсальные и робастные признаки



empirical evaluation: PCA of patch features

empirical evaluation: Patch matching

(Vehicles) (Birds / Airplanes)



(Elephants) (Drawings / Animals)

DINOv2: Learning Robust Visual Features without Supervision. Meta AI Research, **2024**.

Уровень признаков

Image Captioners Are Scalable Vision Learners Too

Pretraining: Какая прокси-задача лучше всего? (Аналог классификации на ImageNet)

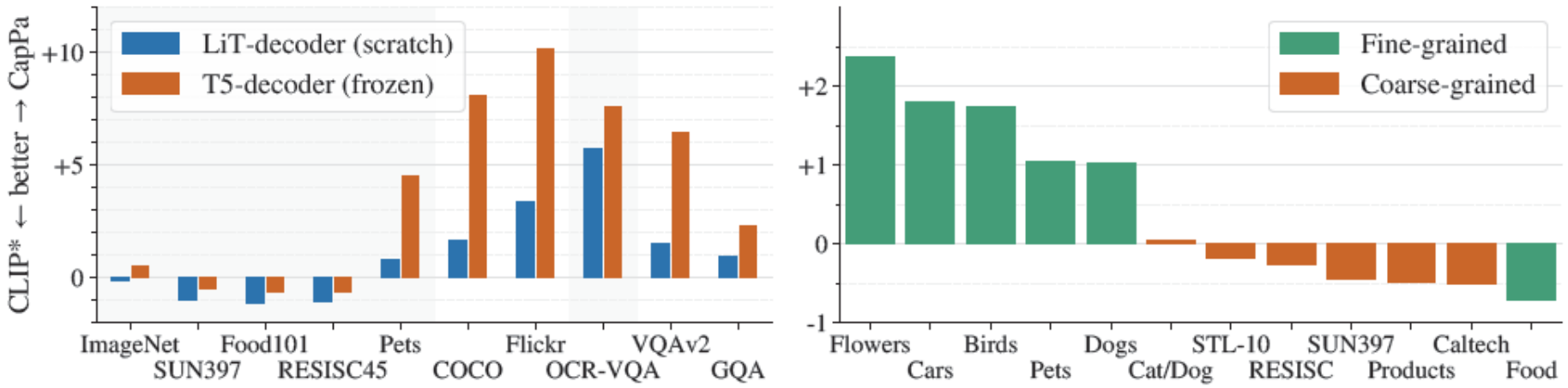


Figure 4: Absolute improvement of CapPa over CLIP* in various settings. **Left:** CapPa pairs significantly better with decoders in image-language tasks, especially when the decoder is a pre-trained and frozen language model. **Right:** CapPa seems to be a noticeably better frozen feature extractor for fine-grained classification tasks (we show L/14 results, see appendix for B/16).

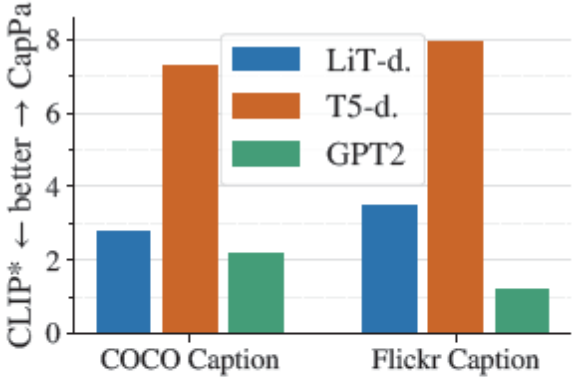


Figure 5: Absolute improvement (single-task) of CapPa over CLIP* for a decoder trained from scratch (LiT-d.), a frozen T5 decoder, and a frozen GPT-2 similar to [44].

We presented an extensive comparison of vision encoders pretrained with a contrastive and generative (captioning) objective and found that the **generatively pretrained encoders obtain better performance when used for captioning, VQA, fine-grained and few-shot classification tasks**, while achieving competitive performance in classification overall. In conclusion, we established **plain image captioning as a competitive pretraining strategy for vision backbones from image-text data**

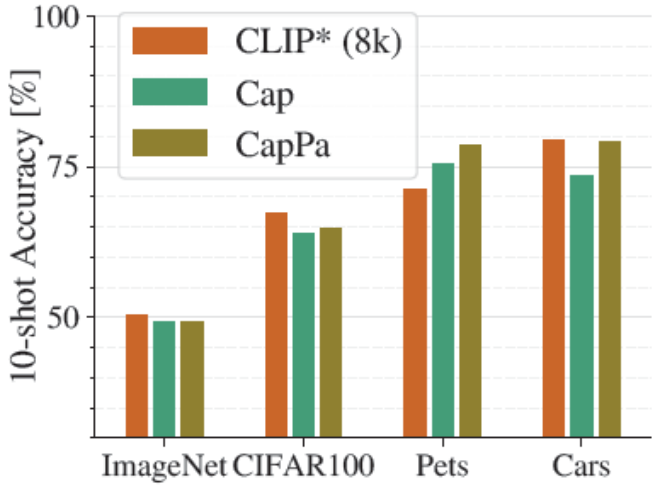
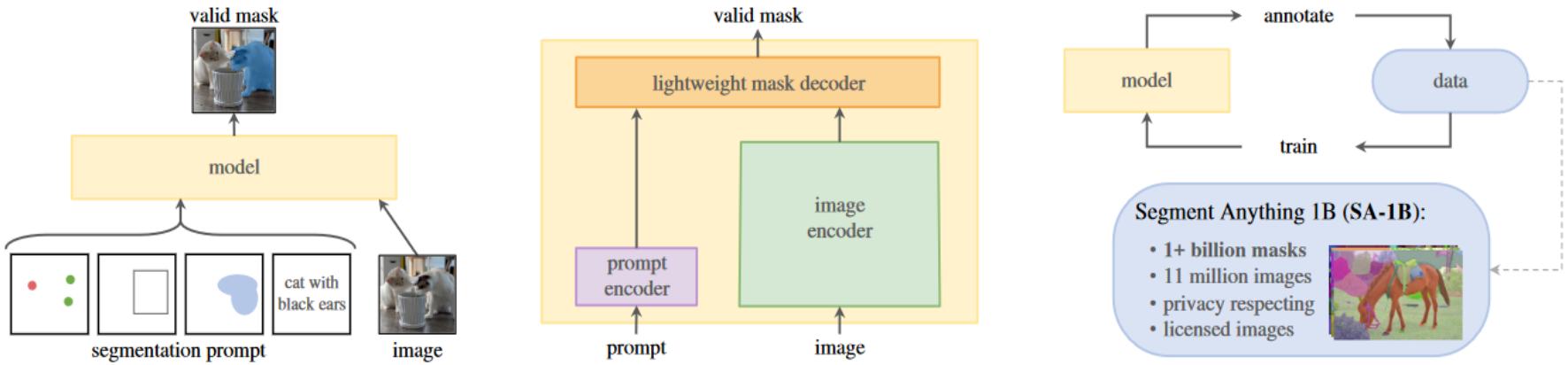


Figure 7: Pretraining on LAION-400M.

Image Captioners Are Scalable Vision Learners Too. Google DeepMind, **2023.**

SAM: Segment Anything, Foundation model for segmentation

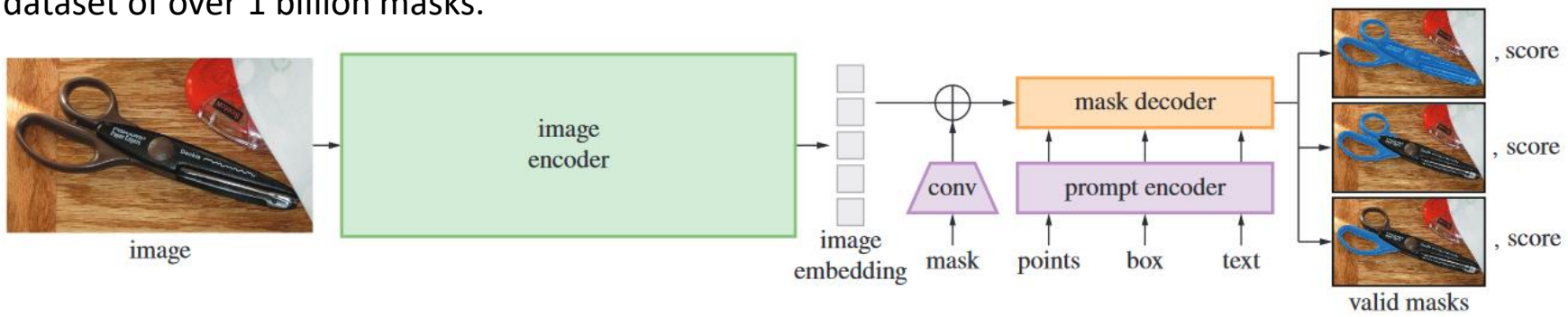
Задача сегментации столь же фундаментальна!



(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Figure 3: Each column shows 3 valid masks generated by SAM from a single ambiguous point prompt (green circle).

Foundation model for segmentation with three interconnected components: a promptable segmentation task, a segmentation model (SAM) that powers data annotation and enables **zero-shot transfer to a range of tasks via prompt engineering**, and a data engine for collecting SA-1B, our dataset of over 1 billion masks.



Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

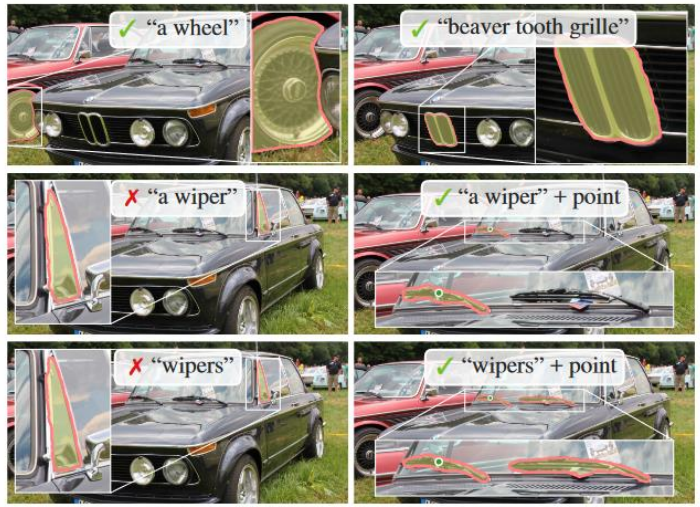


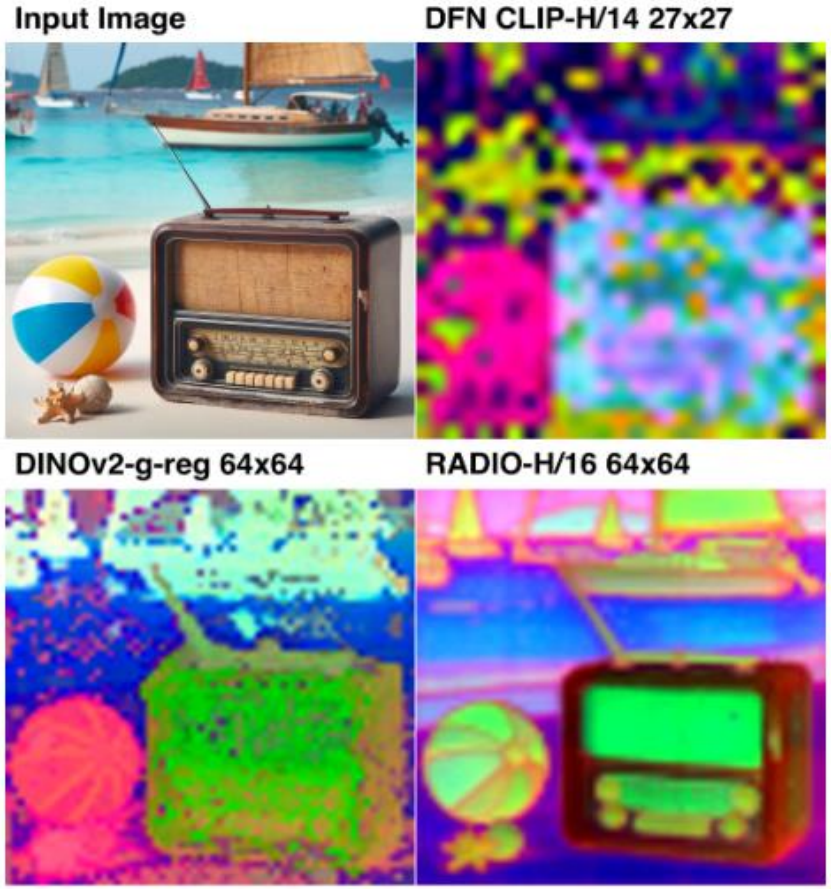
Figure 12: Zero-shot text-to-mask. SAM can work with simple and nuanced text prompts. When SAM fails to make a correct prediction, an additional point prompt can help.

Segment Anything, Alexander Kirillov, …, Piotr Doll´ar, Ross Girshick, Meta AI Research, FAIR, 2023

AM-RADIO (апрель 2024): Vision Foundation Model for All Domains

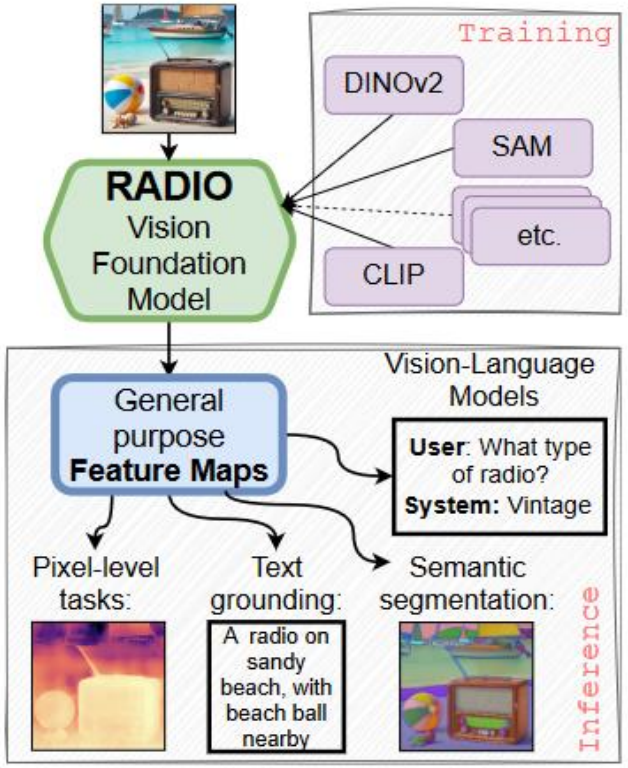
Все уровни, все признаки!

PCA feature visualization of different models

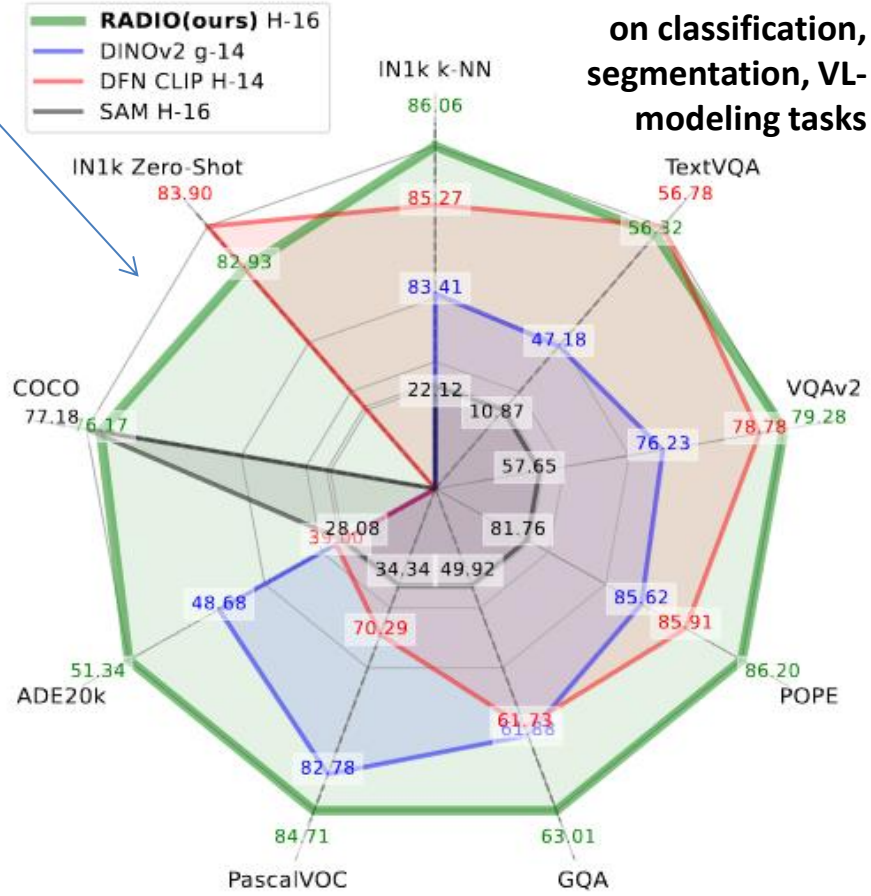


Универсальная фундаментальная модель для задач CV!

the overview of the AM-RADIO framework



benchmarks on classification, segmentation, VL-modeling tasks



AM-RADIO is a framework to distill multiple pretrained vision foundation models, such as CLIP, DINOv2, SAM, into a single model that we call RADIO. As a result, a single vision foundation model agglomerates unique properties of the original models. This unifying approach obtains state-of-the-art feature representations in a single forward pass while also enabling unique properties such as **zero-shot classification (CLIP) or open set instance segmentation (SAM) at negligible additional cost**. Our proposed RADIO model can **process any resolution and aspect ratio, and produces semantically rich dense encodings**.

AM-RADIO: Agglomerative Vision Foundation Model Reduce All Domains Into One, NVIDIA, 2024

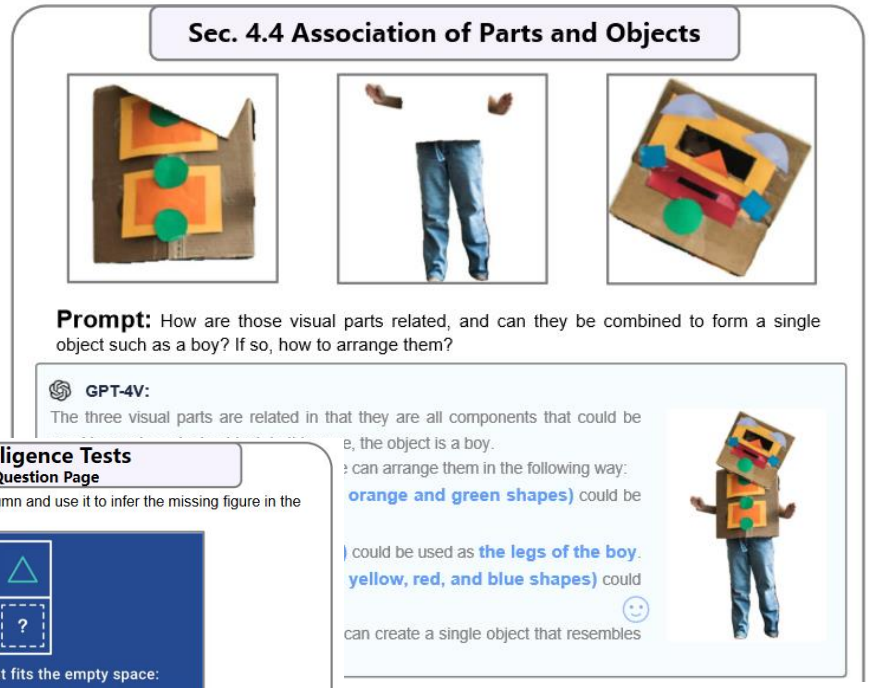
Multimodal Large Language Models

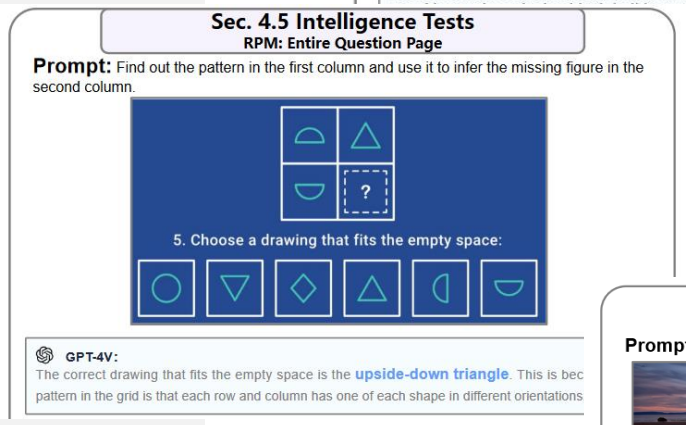
(другой уровень VL-задач)

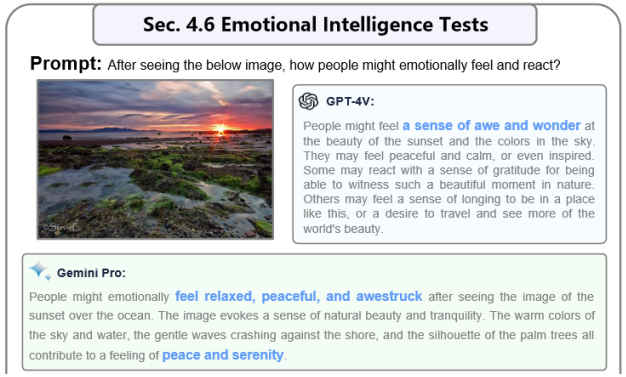
Уровни анализа в VL моделях

Уровень модели мира	**когда нужно не только видеть сцену, но и понимать мир**
Уровень сцены	**классификация и описание сцен (отношения, действия), VQA**
Уровень объектов	**обнаружение объектов, семантическая сегментация…**
Уровень признаков	**особые точки, линии, углы, области, контуры, текстура…**
Уровень изображения	**пикселы (положение, яркость, цвет)**

Gemini vs GPT-4V: Comparison of Vision-Language Models

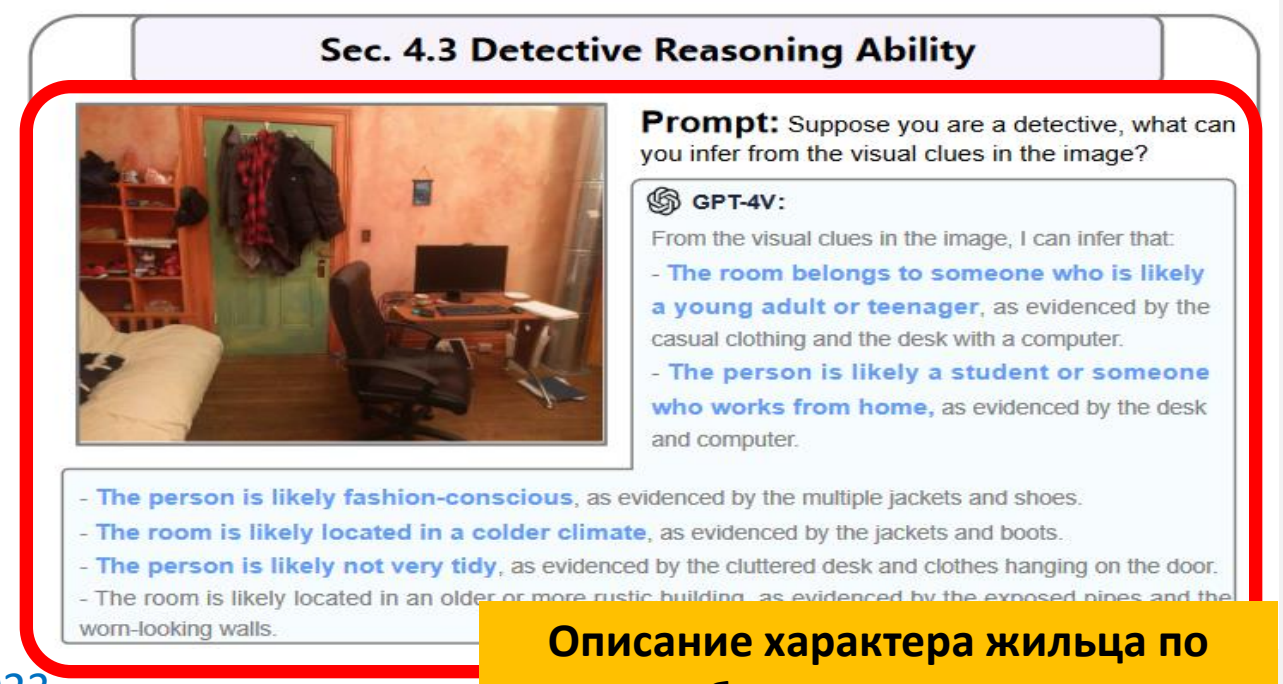






Multimodal evaluation

- **Image Recognition and Understanding:** Sec. 2 addresses the fundamental recognition and comprehension of image content without involving further inference, including tasks such as identifying landmarks, foods, logos, abstract images, autonomous driving scenes, misinformation detection, spotting differences, and object counting.
- **Text Recognition and Understanding in Images:** Sec. 3 concentrates on text recognition (including OCR) within images, such as scene text, mathematical formulas, and chart & table text recognition. Similarly, no further inference of text content is performed here.
- **Image Inference Abilities:** Beyond basic image recognition, Sec. 4 involves more advanced reasoning. This includes understanding humor and scientific concepts, as well as logical reasoning abilities like detective work, image combinations, look for patterns in intelligence tests (IQ Tests), and emotional understanding and expression (EQ Tests).



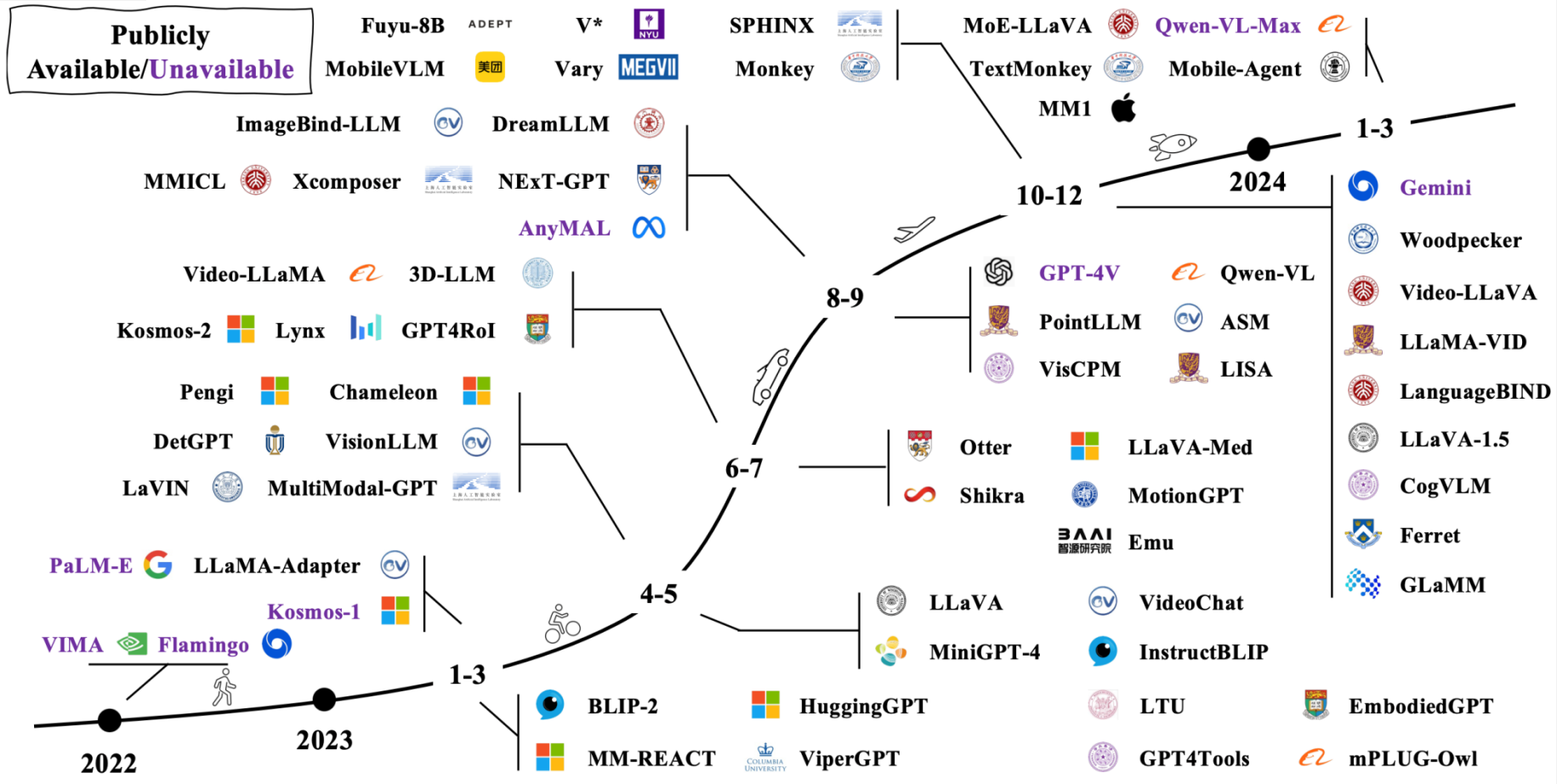
Описание характера жильца по изображению комнаты

OpenAI. Gpt-4v(ision) system card. Technical report, OpenAI, 2023.

Google Gemini Team. Gemini: A family of highly capable multimodal models. 2023.

Gemini vs GPT-4V : A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases, 2023

MLLM: Awesome-Multimodal-Large-Language-Models



- Awesome Papers
  - Multimodal Instruction Tuning
  - Multimodal In-Context Learning
  - Multimodal Chain-of-Thought
  - LLM-Aided Visual Reasoning
  - Foundation Models
  - Evaluation
  - Multimodal Hallucination
  - Multimodal RLHF
  - Others
- Awesome Datasets
  - Datasets of Pre-Training for Alignment
  - Datasets of Multimodal Instruction Tuning
  - Datasets of In-Context Learning
  - Datasets of Multimodal Chain-of-Thought
  - Datasets of Multimodal RLHF
  - Benchmarks for Evaluation
  - Others

Latest Papers and Datasets on Multimodal Large Language Models, and Their Evaluation

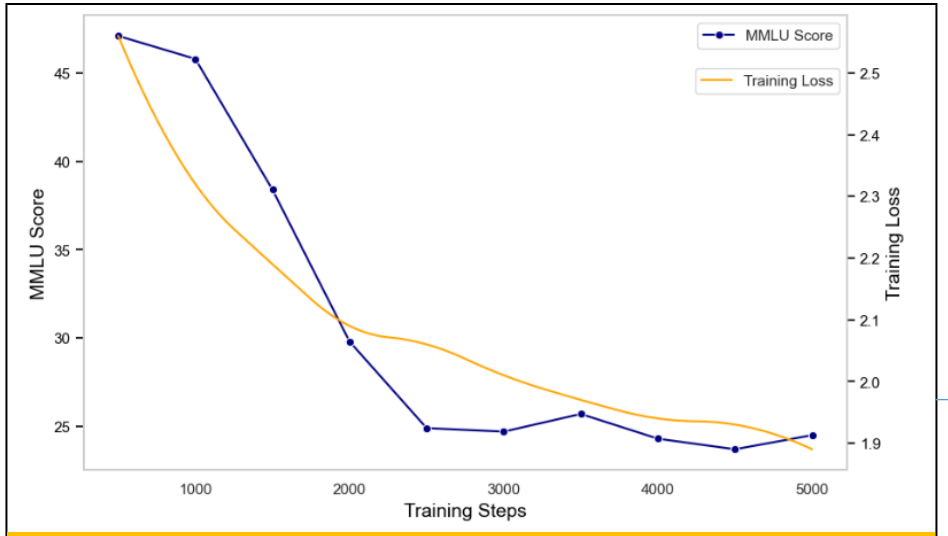
https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

A Survey on Multimodal Large Language Models, 2024

MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, 2024

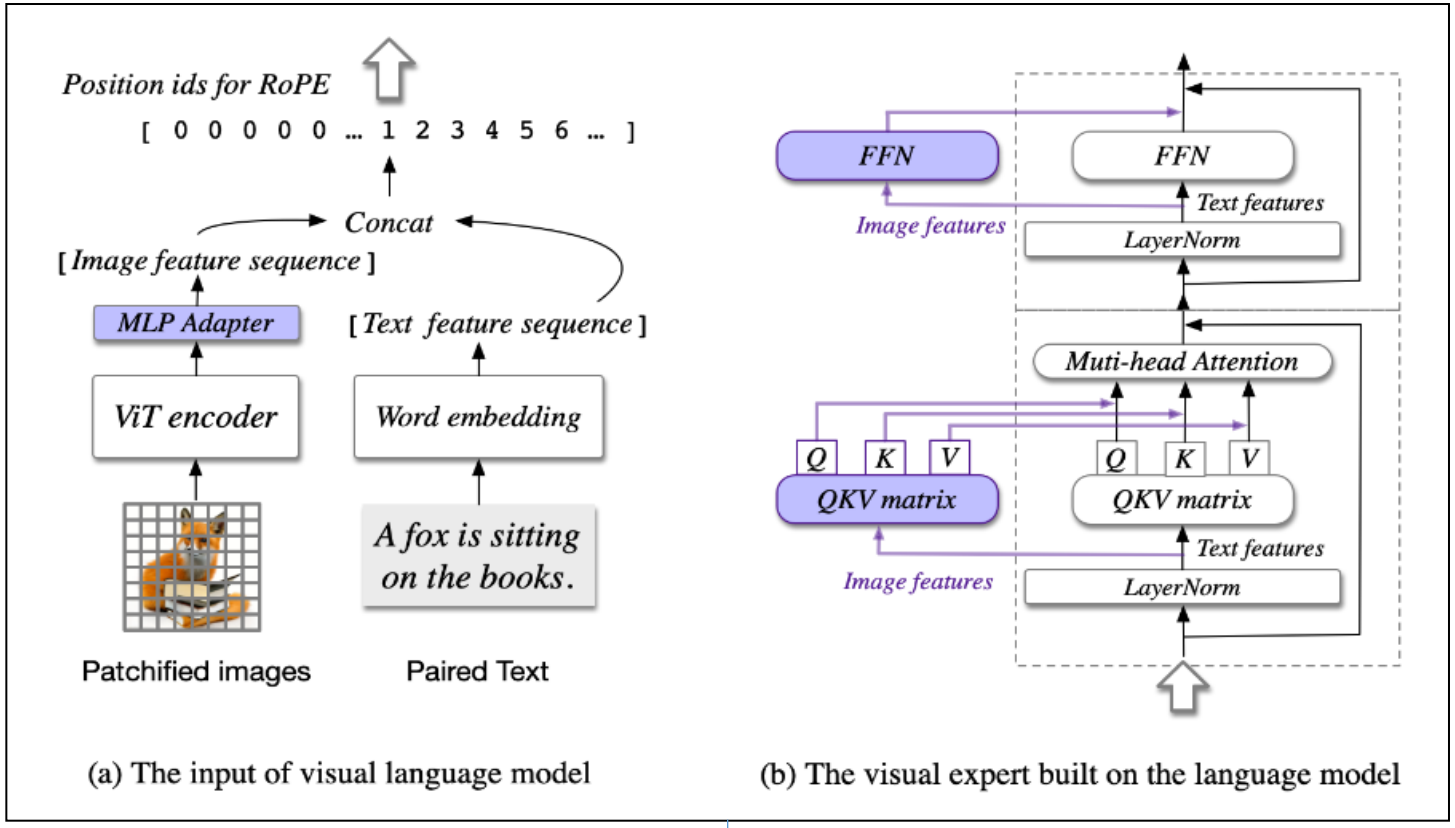
Теперь есть одна страница, где стоит читать про MLLM!

CogVLM: Vision-Language *Deep Alignment*



Как избежать катастрофического забывания базовых модальностей при мультимодальном обучении?

VLM using the LAION dataset, the model's score on the pure text dataset MMLU rapidly decreases, dropping to 24.9 at 2500 steps.



(a) The input of visual language model

(b) The visual expert built on the language model

Figure 4. **The architecture of CogVLM.** (a) The illustration about the input, where an image is processed by a pretrained ViT and mapped into the same space as the text features. (b) The Transformer block in the language model. The image features have a

Гипотеза: для того, чтобы показывать визуальное понимание уровня GPT-4V, нужно правильно объединять предобученные модели разной модальности (LLM и ViT)

Добавим к ViT адаптер и обучим визуального эксперта к обученной LLM для задач NLP

CogVLM: Visual Expert for Pretrained Language Models, 2024

CogVLM: Vision-Language *Deep Alignment*

Visual Reasoning

Q: From the perspective of a photographer, analyze why this picture was taken well?

CogVLM: From a photographer's perspective, several elements contribute to the quality of this image. The composition is balanced, with the hot air balloon taking up a significant portion of the frame, yet still allowing the surrounding environment to be visible. The use of lighting, especially the soft glow from the balloon and the car's headlights, creates a warm and inviting atmosphere. The depth of field is shallow, focusing on the subjects in the foreground while softly blurring the background, drawing attention to the main elements of the scene.

Grounding with Caption

Q: Can you provide a description of the image and include the coordinates [[x0,y0,x1,y1]] for each mentioned object?

CogVLM: A piggy bank [[302,198,694,857]] with a smile [[400,520,582,643]] on it is surrounded by coins [[028,820,986,997]], toy houses [[052,697,174,892; 526,763,636,952;168,697,286,883;738,697,852,895]], and greenery.

Grounding Visual Question Answering

Q1: What color are the clothes of the girl whose hands are holding flowers? Let's think step by step.

CogVLM: Find the flowers [[328,519,436,780]]. Check the girl holding it, got the girl [[116,069,450,927]]. Check the clothes of it, got the clothes [[134,309,418,852]]. The clothes [[134,309,418,852]] are blue. So the answer is blue.

Q2: What color are the jeans of the girl whose hands are not holding flowers? Let's think step by step.

CogVLM: Find the not(flowers) hands [[654,637,710,728]]. Check the girl with it, got the girl [[526,109,880,946]]. Check the jeans of it, got the jeans [[530,613,880,871]]. The jeans [[530,613,880,871]] are gray. So the answer is gray.

Вывод: для визуального понимания уровня GPT-4V не нужна Very Large LM! *Важнее правильно объединять предобученные визуальные и текстовые модели (см. PEFT, адаптеры)*

CogVLM: Visual Expert for Pretrained Language Models, 2024

Сегодня кратко

Модели и методы многомодального синтеза данных

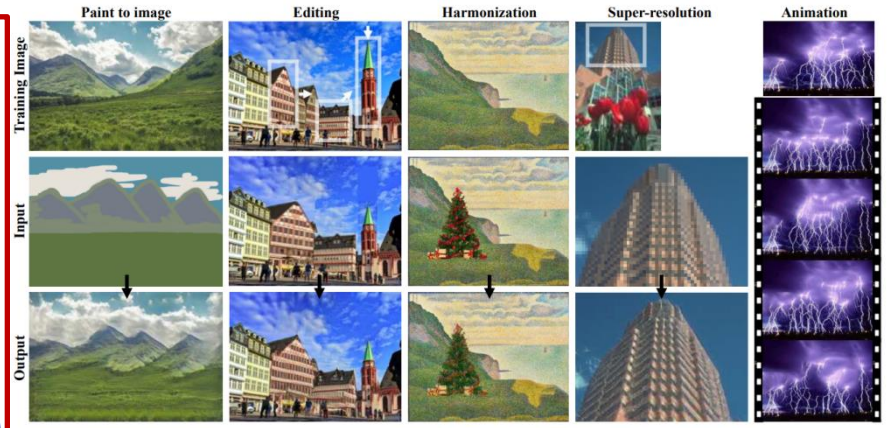
Генеративные модели

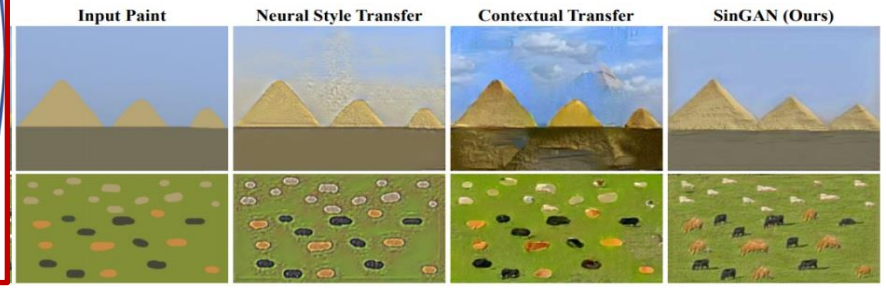
Модели для генерации несимвольных данных

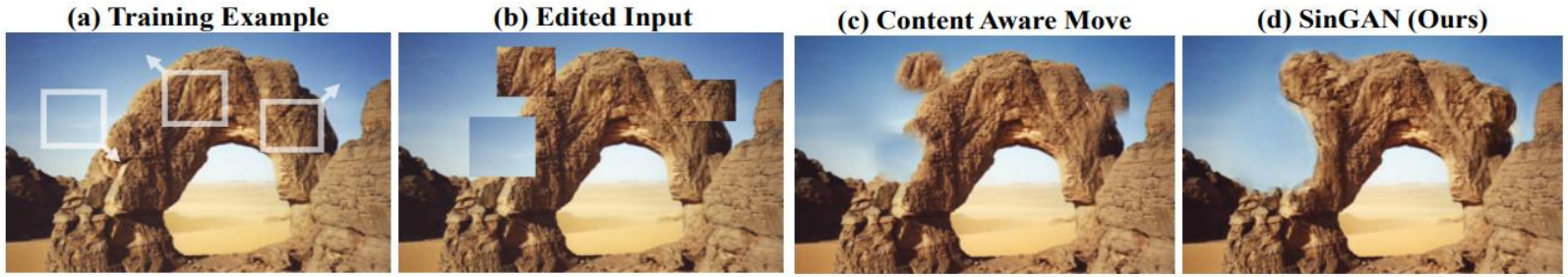
Generative Adversarial Networks (GAN)

GM до 2020

- GAN
  - Realistic samples
  - Suffer from mode-collapse
- Normalizing flow
  - Explicit model
  - Tend to overfit training data
  - Not flexible enough
- VAE
  - Covers training sample
  - Unrealistic samples (e.g. blurry images)







Dmitry Vetrov, OpenTalks.AI/2022

Adaptive Deep Models across Visual Domains, Kate Saenko, ICCV, 2017

SinGAN: Learning a Generative Model From a Single Natural Image, Shaham et. al, 2019

Проблемы и вызовы в GM (2020)

Слабая генерализация за пределами выборки

Нет генерации образов по словесному описанию

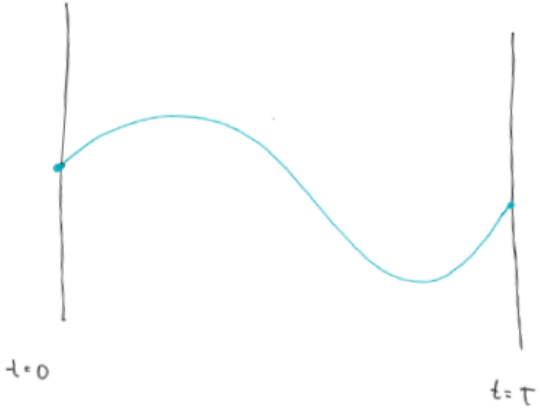
Diffusion models: новый лидер 2020+

References

- J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239, June 2020.
- Y. Song, J. Sohl-Dickstein, D. Kingma, A. Kumar, S. Ermon, B. Poole. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020.

- Ordinary differential equation defines the evolution of a point

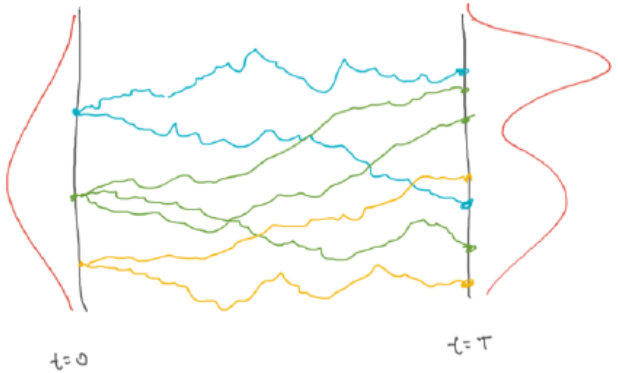
$$dx = f(x,t)dt$$



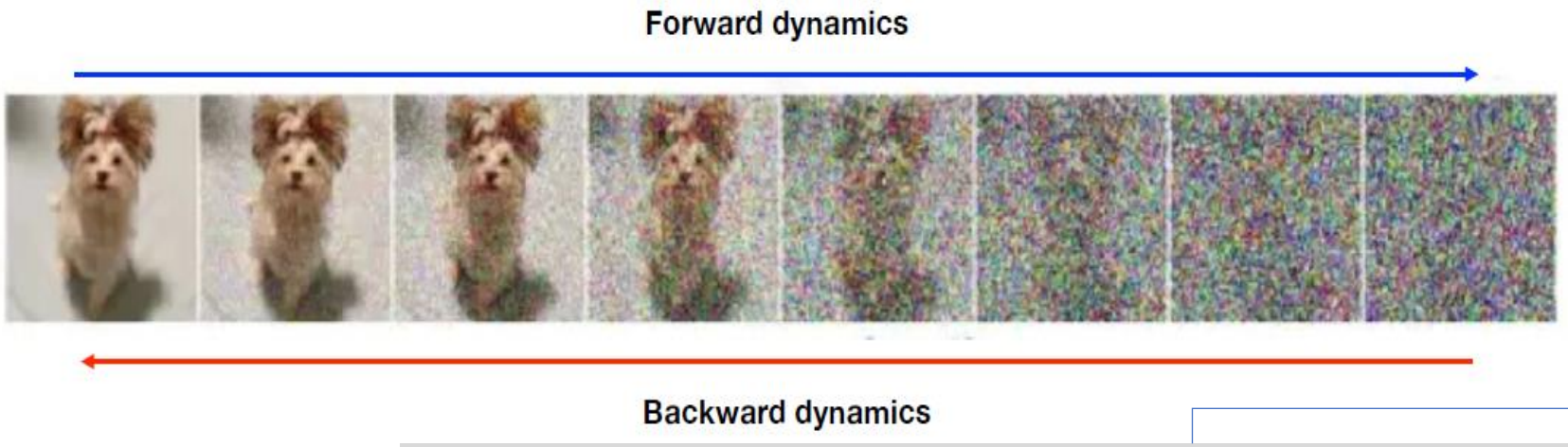
Модель 1: решение обратного стохастического ДУ

- Stochastic differential equation defines the evolution of a distribution

$$dx = f(x,t)dt + g(x,t)dW$$



Задача: выучить распределение образов для класса на примерах, а потом сэмплировать из этого распределения

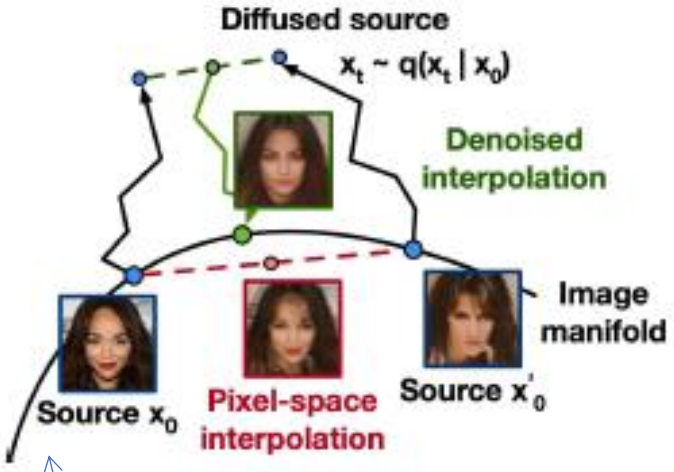


Модель 2: поэтапное восстановление зашумленного образа

Кстати, это и есть «усилитель смысла» Эшби, извлекающий смысл из шума!

Denoising diffusion probabilistic models

Интерполяция между двумя лицами: в каждой точке только лица!



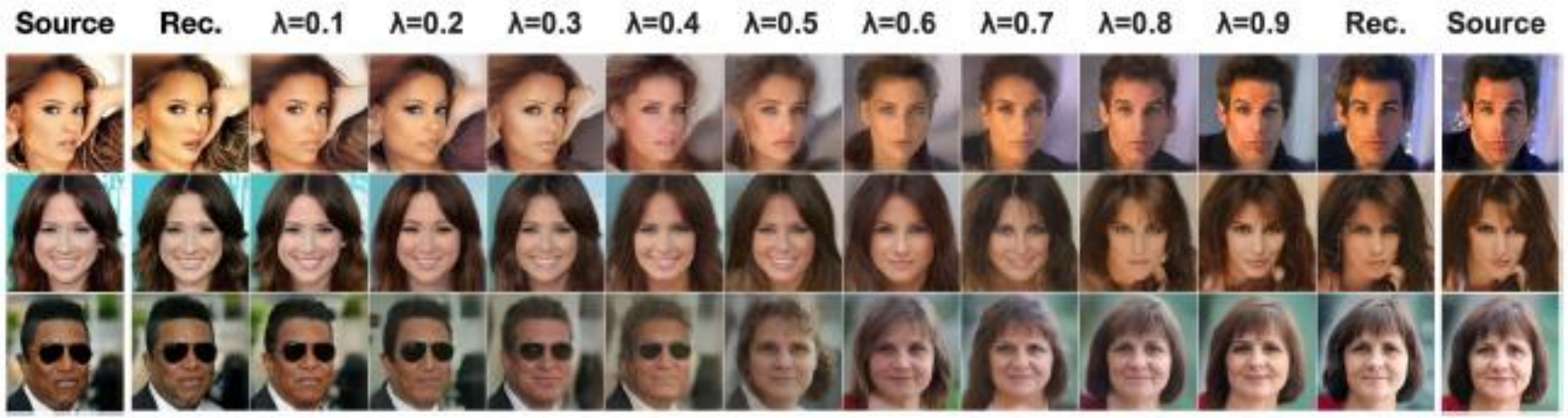
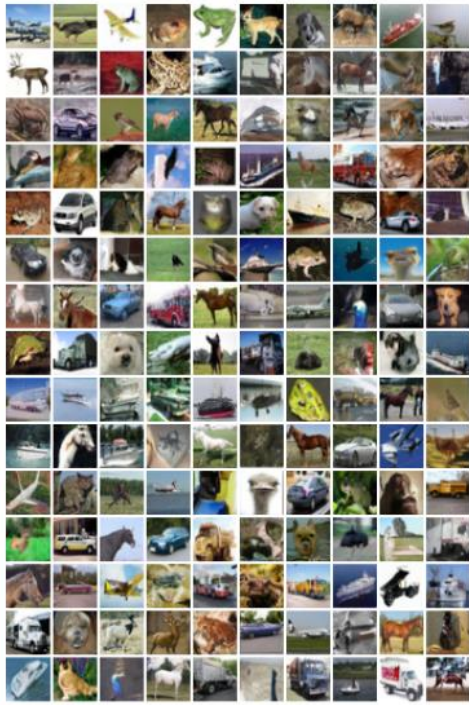


Figure 8: Interpolations of CelebA-HQ 256x256

Непрерывное пространство (распределение) изображений

Выучили распределение для лиц

Фото людей, которых никогда не существовало



J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. NeurIPS 2020

Text-2-Image, Text-2-Video

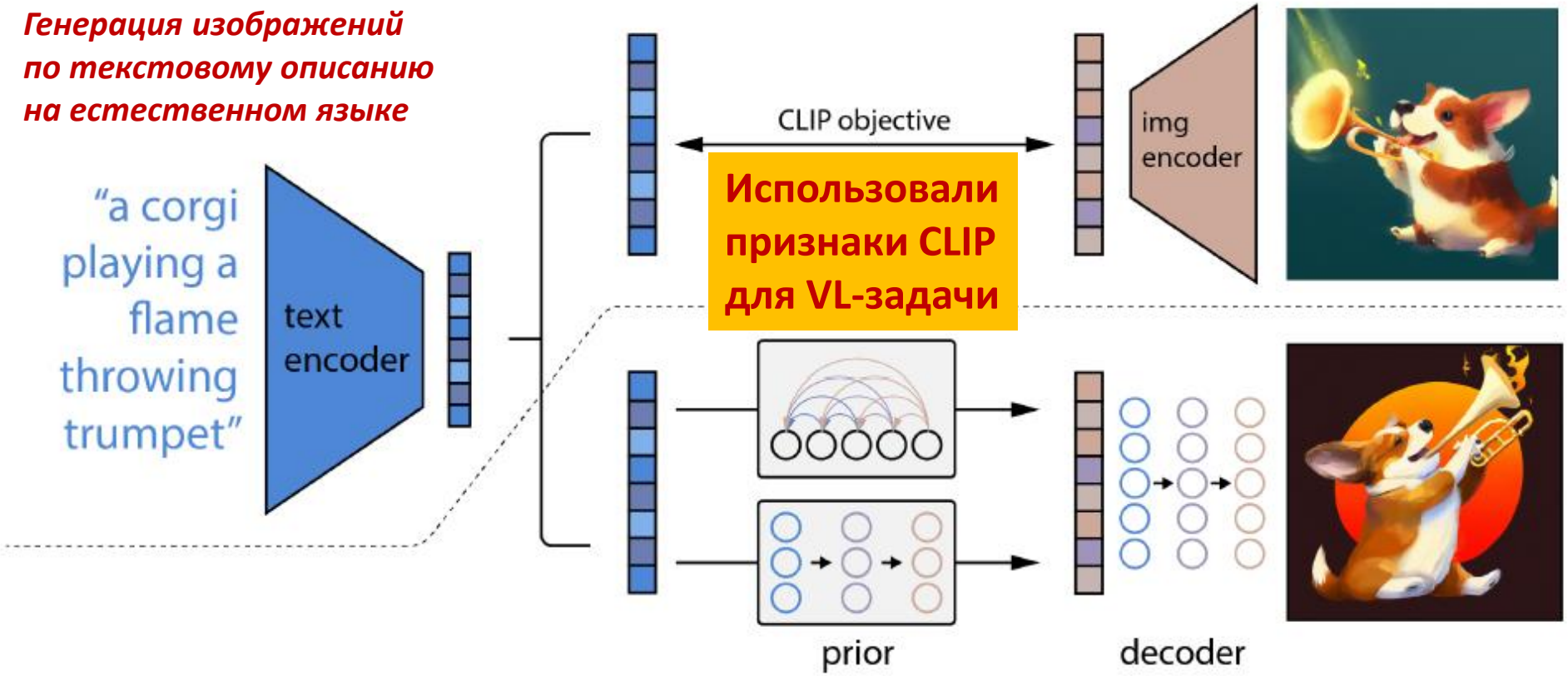
(генерация изображений и видео по словесному описанию)

DALL·E 2: images and art from natural language description

Задача: text-to-image generation

Генерация изображений по текстовому описанию на естественном языке

Соединение 2 типов генеративных моделей: **трансформер** для текстовой модальности, **диффузная модель** для изображения



Использовали признаки CLIP для VL-задачи

A high-level overview of unCLIP. Above the dotted line, we depict the CLIP training process, through which we learn a joint representation space for text and images. Below the dotted line, we depict our text-to-image generation process: a CLIP text embedding is first fed to an autoregressive or **diffusion prior** to produce an image embedding, and then this embedding is used to condition a **diffusion decoder** which produces a final image. Note that the CLIP model is frozen during training of the prior and decoder.

Hierarchical Text-Conditional Image Generation with CLIP Latents
Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

We develop methods for training **diffusion priors in latent space**, and show that they achieve comparable performance to autoregressive priors, while being more compute-efficient.

DALL·E 2: images and art from natural language description

Непрерывное пространство изображений (распределение) стало семантическим!

Пространство признаков CLIP

a photo of a cat → an anime drawing of a super saiyan cat, artstation

a photo of a victorian house → a photo of a modern house

a photo of an adult lion → a photo of lion cub

a photo of a landscape in winter → a photo of a landscape in fall

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen, OpenAI/2022

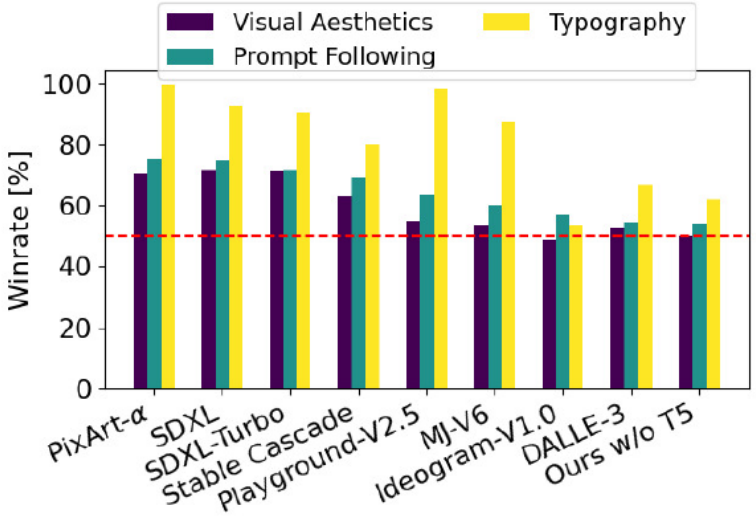
diffusion priors in latent space

Stable Diffusion 3: текущий лидер (2024)

a space elevator, cinematic scifi art | A cheeseburger with juicy beef patties and melted cheese sits on top of a toilet that looks like a throne and stands in the middle of the royal chamber. | a hole in the floor of my bathroom with small gremlins living in it | a small office made out of car parts | This dreamlike digital art captures a vibrant, kaleidoscopic bird in a lush rainforest. | human life depicted entirely out of fractals | an origami pig on fire in the middle of a dark room with a pentagram on the floor

an old rusted robot wearing pants and a jacket riding skis in a supermarket.

smiling cartoon dog sits at a table, coffee mug on hand, as a room goes up in flames. "This is fine," the dog assures himself.



Human Preference Evaluation against current closed and open SOTA generative image models. Our 8B model compares favorable against current state-of-the-art text-to-image models when evaluated on the parti-prompts (Yu et al., 2022) across the categories visual quality, prompt following and typography generation.

Outperforms DALL·E 3, Midjourney v6, and Ideogram v1

Our new **Multimodal Diffusion Transformer (MMDiT)** architecture **uses separate sets of weights for image and language representations**, which improves text understanding and spelling capabilities compared to previous Stable Diffusion.

Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. 2024.

ControlNet: Adding Conditional Control to Text-to-Image Diffusion Models

Figure 1: Controlling Stable Diffusion with learned conditions. ControlNet allows users to add conditions like Canny edges (top), human pose (bottom), *etc.*, to control the image generation of large pretrained diffusion models. The default results use the prompt "a high-quality, detailed, and professional image". Users can optionally give prompts like the "chef in kitchen".

Text-2-Image: ***Добавим к текстовому запросу визуальный скетч!***

Задача: управлять генерацией изображения не только словесно

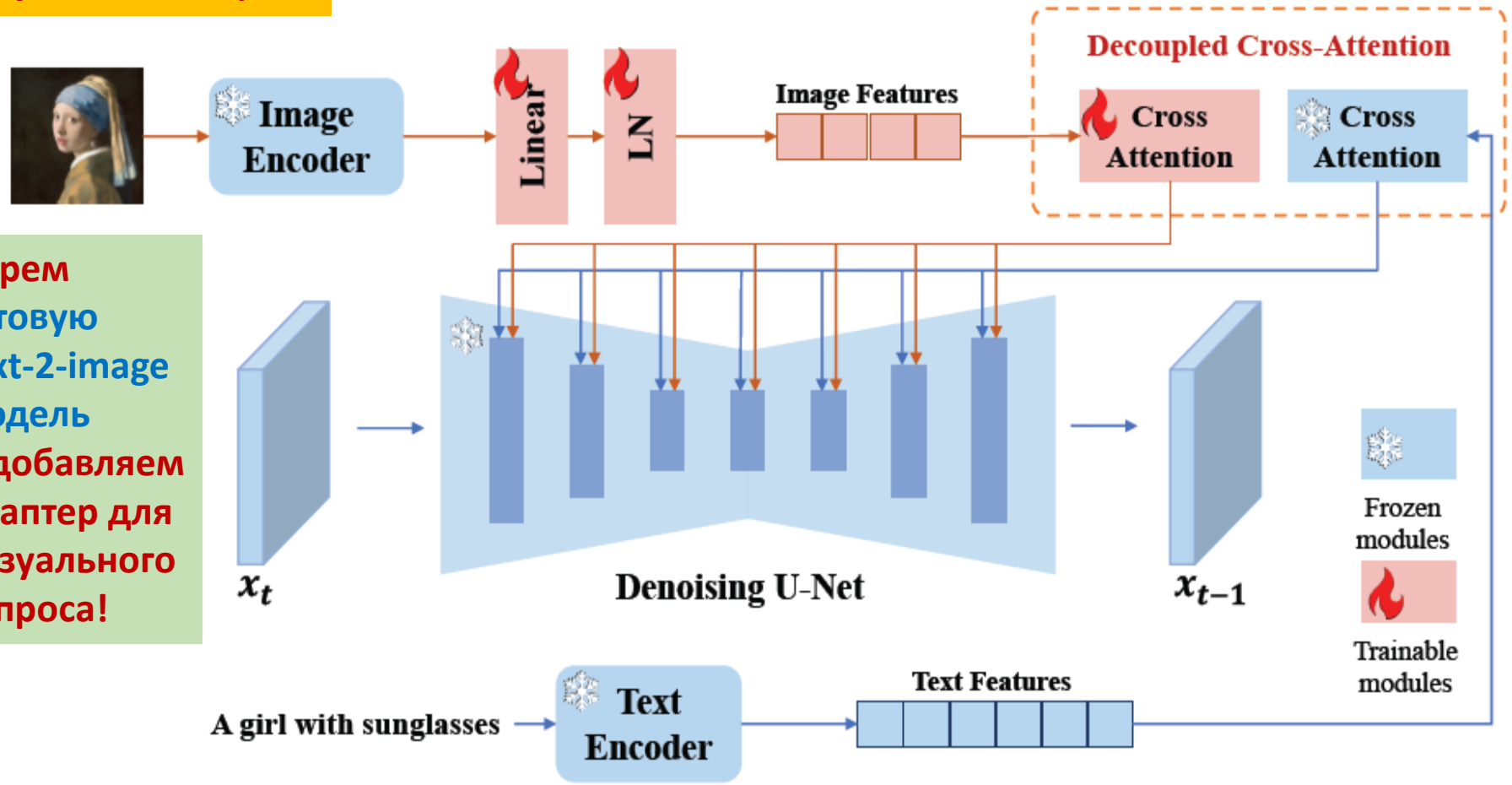
Adding Conditional Control to Text-to-Image Diffusion Models. 2023.

IP-Adapter: Text Compatible Image Prompt Adapter

Text-2-Image:
Не добавим что-то к тексту, а подадим сразу многомодальный текстовый + визуальный запрос!

The overall architecture of our proposed IP-Adapter with decoupled cross-attention strategy. Only the newly added modules (in red color) are trained while the pretrained text-to-image model is frozen.

Берем готовую text-2-image модель и добавляем адаптер для визуального запроса!



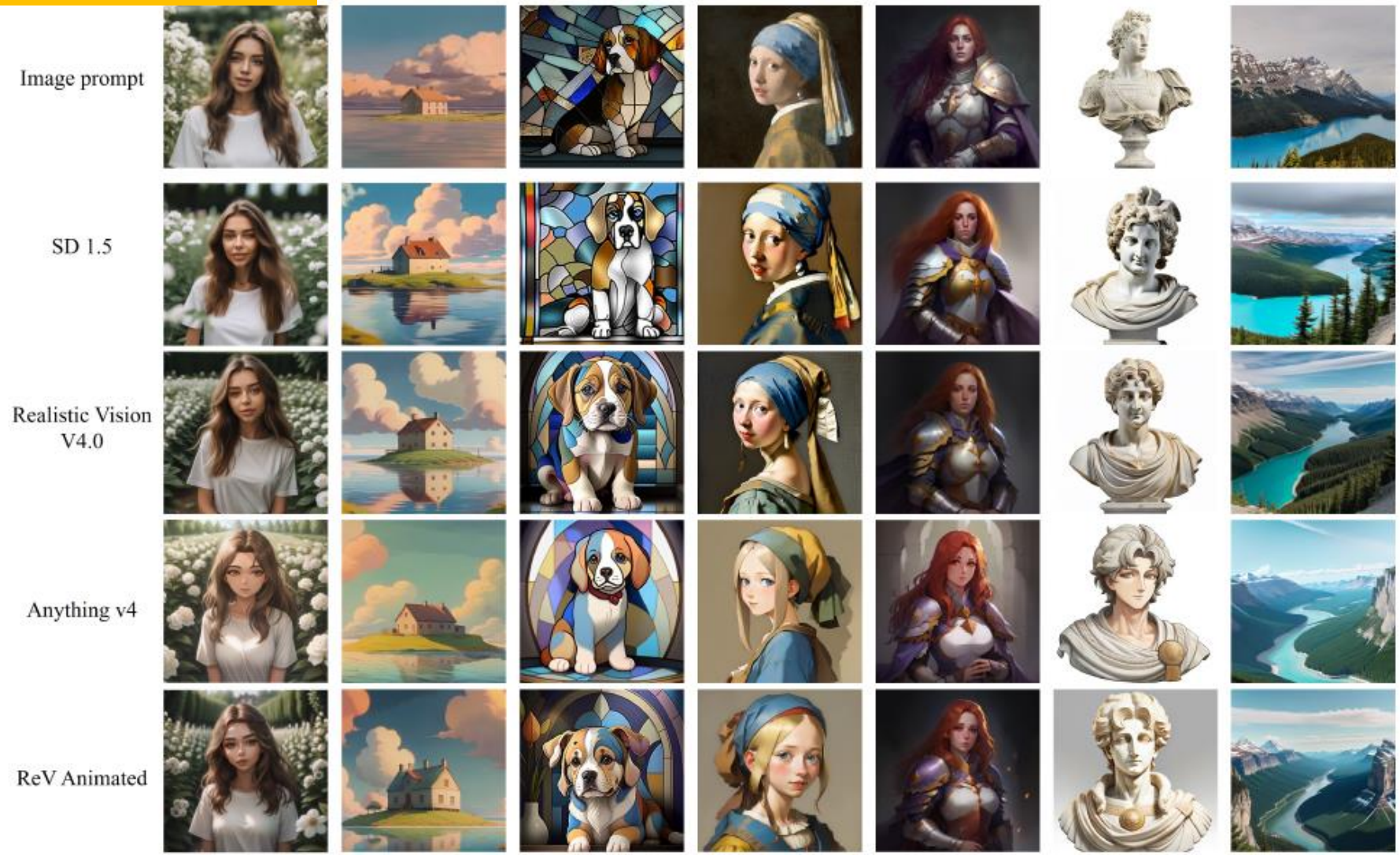
The main problem: cross-attention layers in the pretrained diffusion model are trained to adapt the text features. Consequently, merging image features and text features only accomplishes the alignment of image features to text features, but **potentially misses some image-specific information from the reference image.** Our IP-Adapter adopts a decoupled cross-attention mechanism**. For every cross-attention layer in the UNet of diffusion model, we add an additional cross-attention layer only for image features.** In the training stage, **parameters of the new cross-attention layers are trained**, while the **original UNet model remains frozen**.

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. 2023.

Text-2-Image:
Текстовый и визуальный запрос!

IP-Adapter: Text Compatible Image Prompt Adapter

Images of different diffusion models with our proposed IP-Adapter. The IP-Adapter is only trained once.



We present **IP-Adapter**, an effective and lightweight adapter to achieve image prompt capability for the pretrained text-to-image diffusion models. **The key design of our IP-Adapter is decoupled cross-attention mechanism that separates cross-attention layers for text features and image features**. IP-Adapter with only **22M parameters can achieve comparable or even better performance to a fully fine-tuned image prompt model**. As we freeze the pretrained diffusion model, the proposed IP-Adapter can be generalized not only to other custom models fine-tuned from the same base model, but also to controllable generation using existing controllable tools

IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models. 2023.

IMAGEN VIDEO: video generation with diffusion models

Melting pistachio ice cream dripping down the cone.

Ролики короткие

A british shorthair jumping over a couch.

Проблема с лишними конечностями

Coffee pouring into a cup.

Камера на движется, 3D образ нарушается

Генерация видео по текстовому описанию

IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS, Google Research, Brain Team, 2022

SORA: Video generation models as world simulators (2024)

Ролики длиннее

https://cdn.openai.com/tmp/s/bike_1.mp4

3D consistency. SORA может создавать видео с динамическим движением камеры. Когда камера сдвигается и вращается, люди и элементы сцены последовательно перемещаются в трехмерном пространстве.

https://openai.com/sora#research

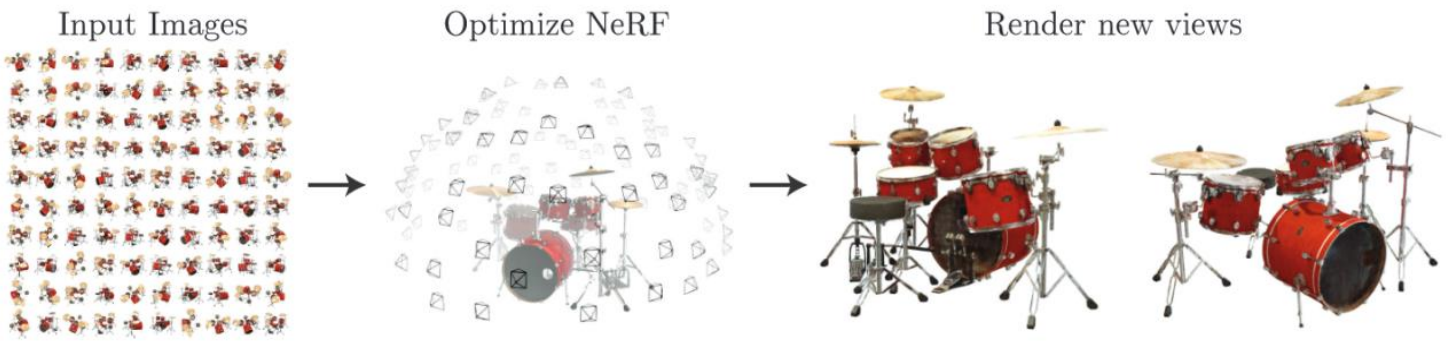
Video generation models as world simulators, OpenAI, 2024

Генерация видео по текстовому описанию

3D Scene Generation & Analysis

Предобученные по ракурсным изображениям описания 3D сцен

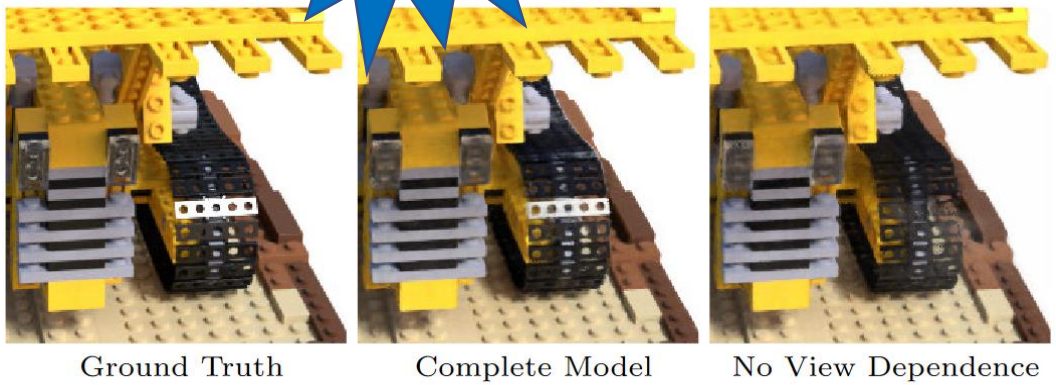
NeRF: Representing Scenes as Neural Radiance Fields

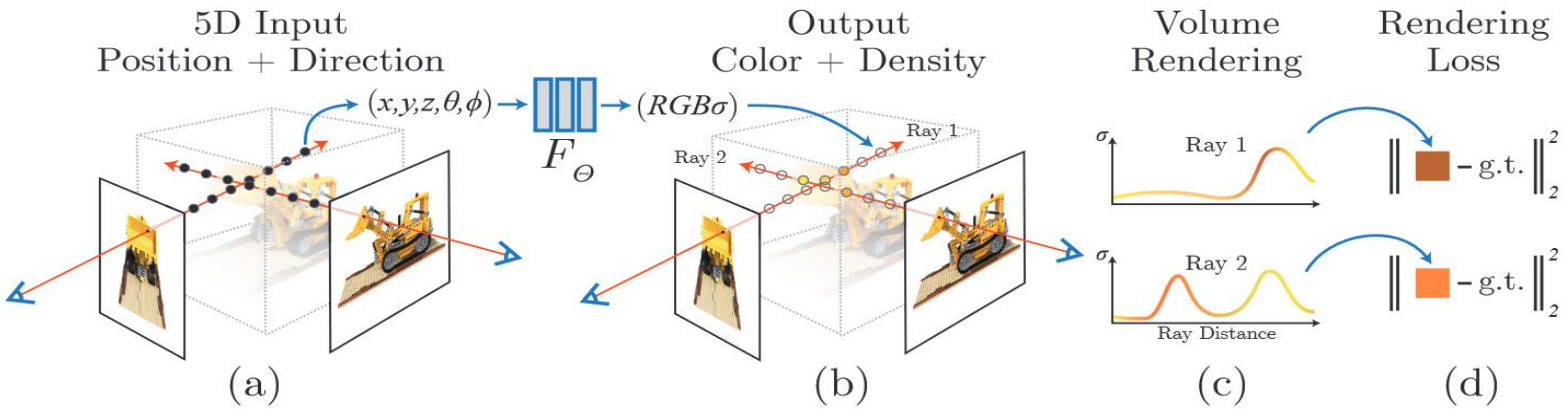


Как соединить традиционные технологии 3D рендеринга с нейронными сетями и машинным обучением

TOP AI 2021!

Continuous 5D neural radiance field representation (volume density and view-dependent color at any continuous location) of a scene from a set of input images. We use techniques from volume rendering to accumulate samples of this scene representation along rays to render the scene from any viewpoint.





- Пропускаем лучи через сцену для семплирования на них множества точек
- Для каждой точки из множества получаем плотность (не зависит от направления взгляда) и цвет (зависит от направления взгляда)
- Применяем рендеринг для создания изображений

Mildenhall et. al, Nerf: Representing scenes as neural radiance fields for view synthesis. UC Berkeley, Google Research, 2021

Развитие и приложения NeRF

Overview

Не только компьютерная графика, но и компьютерное зрение

В исходной задаче генерации изображений 3D сцен и объектов

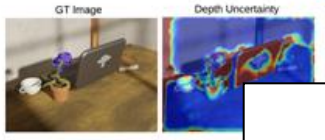
NeRF-RPN: object detection in NeRFs

Object Detection

LERF: Language Embedded Radiance Fields

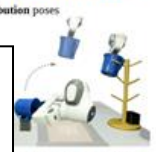
Multi-modal Visual-Language Tasks

NeRF-SLAM: Real-Time Dense Monocular SLAM with Neural Radiance Fields



SLAM: Ориентация роботов и БЛА в пространстве

Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation



Neural Fields for Robotic Object Manipulation from a Single Image

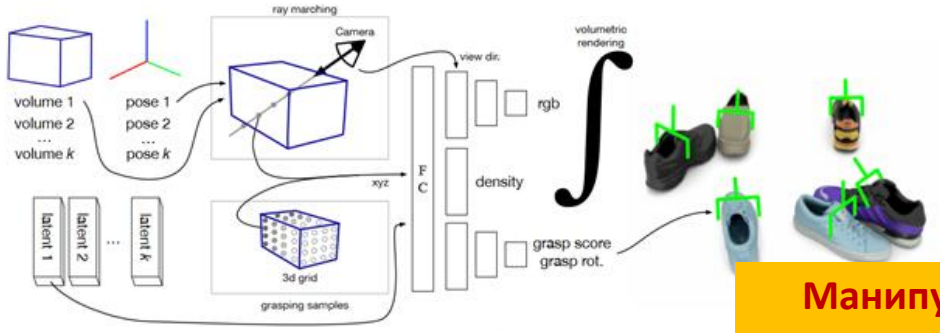


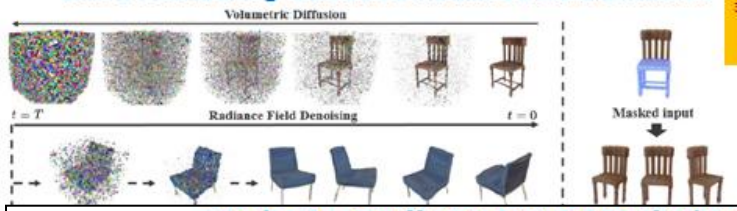
Fig. 1. We present a unified representation that can be used to re-render scenes and generate grasping poses. Given obj left), and learned latent codes (bottom left), the system renders the scene in novel configurations and annotates each o image). The rendering ray-marches through a set of volumes (cuboids in which the objects reside). In order to generate grasps (shown in green on the right), we use a grid sampling approach that covers each object volume. Grasp orientation and score are predicted using a decoder which is partially shared with the volumetric rendering process. We select the grasp point that yields the highest confidence score. FC refers to a fully connected layer.

Манипулирование предметами

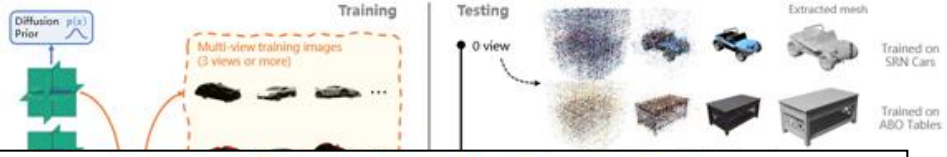
Манипулирование предметами

DiffRF: Rendering-Guided 3D Radiance Field Diffusion

Вернемся к задаче генерации изображений 3D сцен и объектов



Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction



DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models

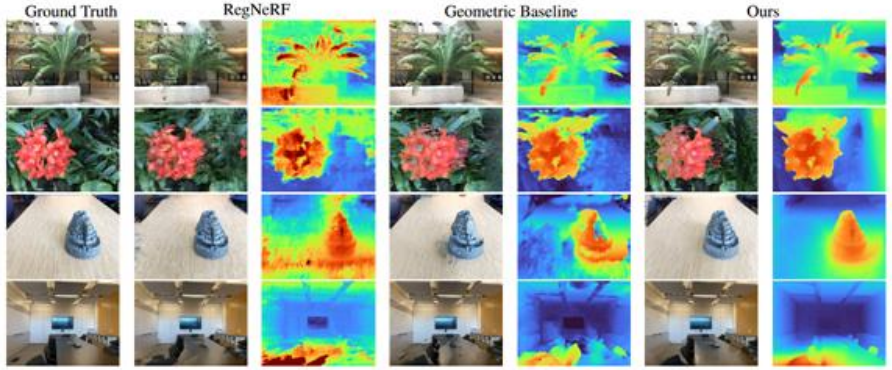
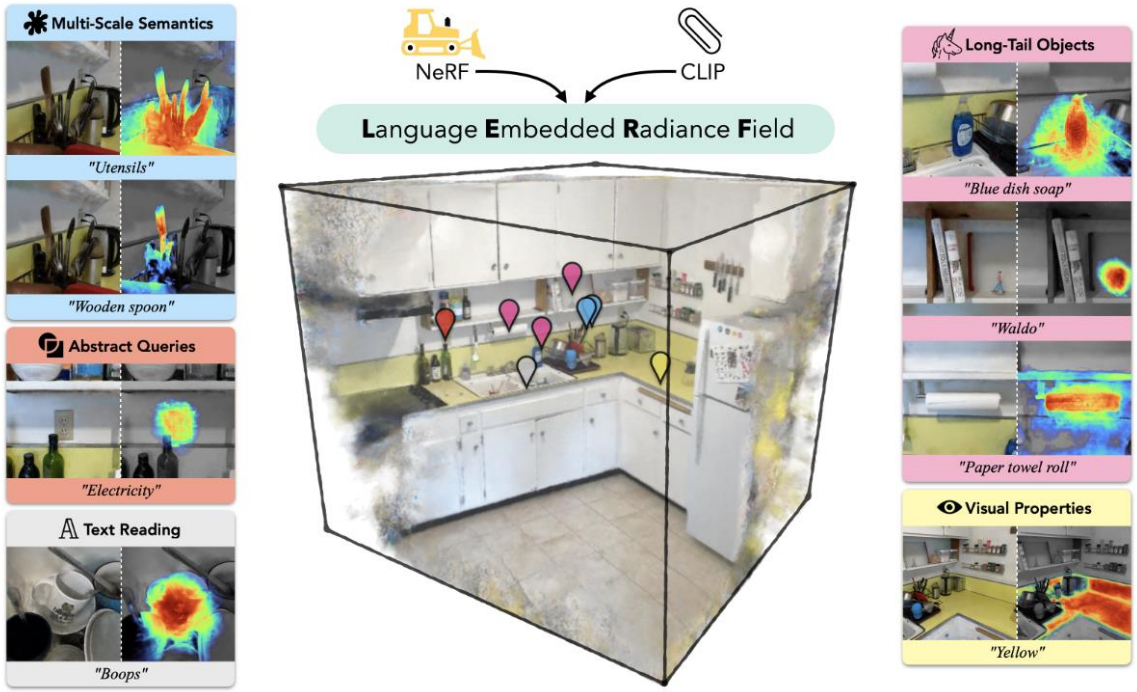


Figure 4. Qualitative results for the task of novel view synthesis on LLFF dataset. NeRF models are trained with 3 views and rendered from one of test views. Our DDM model encourages more realistic geometry as seen in the depth maps.

Соединяем NeRF с диффузными моделями

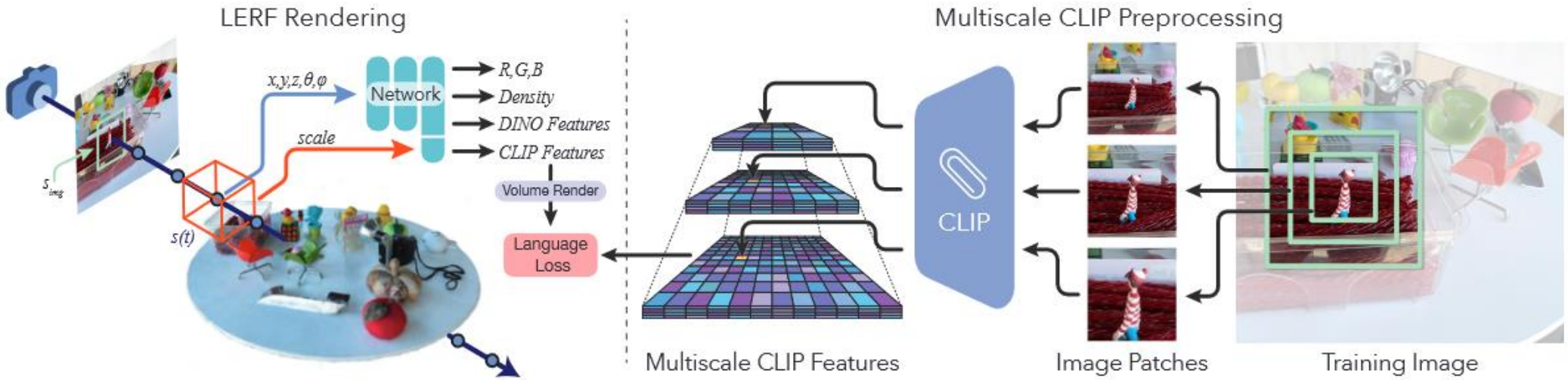
Соединяем NeRF с диффузными моделями

LERF: Language Embedded Radiance Fields



LERF grounds CLIP representations in a dense, multi-scale 3D field. A LERF can be reconstructed from a hand-held phone capture within 45 minutes, then can **render dense relevancy maps given textual queries interactively in real-time**. LERF enables a broad range of concepts to be queried via natural language, from abstract queries like "Electricity", visual properties like "Yellow", long-tail objects such as "Waldo", and even reading text like "Boops" on the mug. For each prompt, an RGB image and relevancy map are rendered focusing on the location with maximum relevancy activation

We pre-compute a multi-scale feature pyramid of CLIP embeddings over training views, and during training interpolate this pyramid with simg and the ray's pixel location to obtain CLIP supervision. The CLIP loss maximizes cosine similarity, and other outputs are supervised with mean squared-error using standard per-pixel rendering.

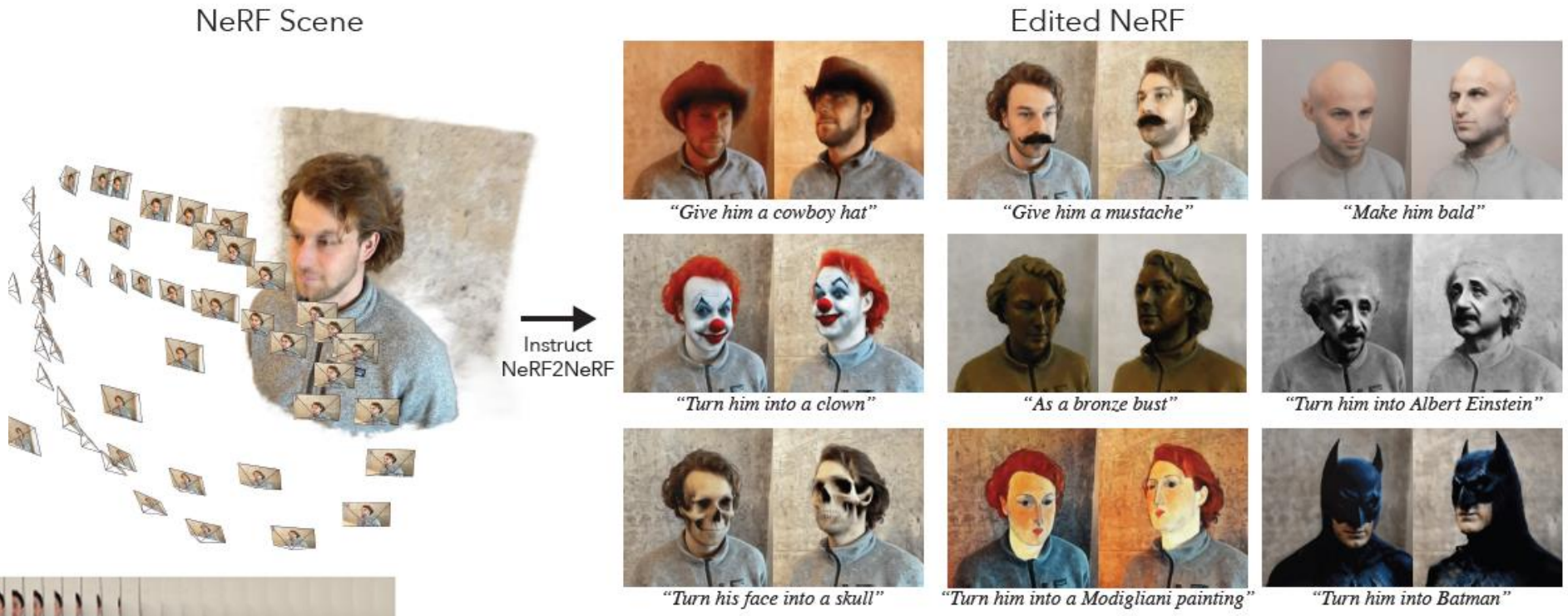


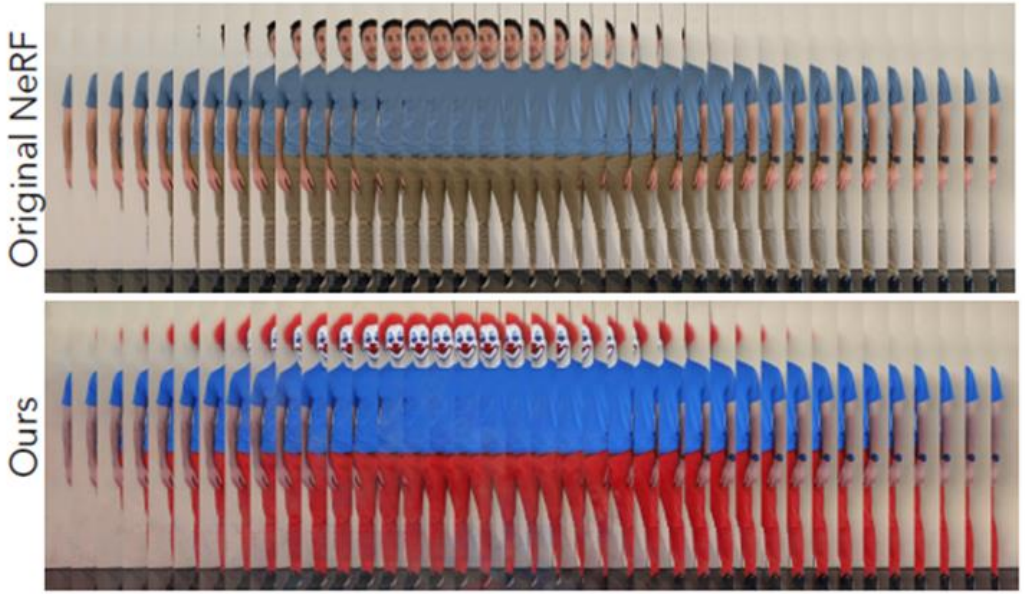
LERF: Language Embedded Radiance Fields, UC Berkeley, 2023

Multi-modal Visual-Language Tasks

Instruct-NeRF2NeRF: вербальное редактирование 3D сцен

Теперь к уже нейросетевой 3D модели NeRF можно добавить различные генеративные возможности нейросетей, в том числе – синтез и редактирование по текстовому описанию…





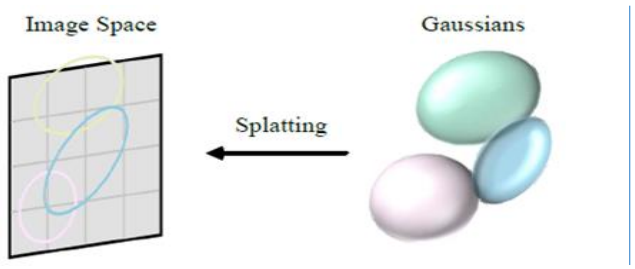
3D consistency. На основе NeRF также можно создавать видео с динамическим движением камеры *в искусственно измененном мире*. Когда камера сдвигается и вращается, измененные элементы сохраняют свой вид.

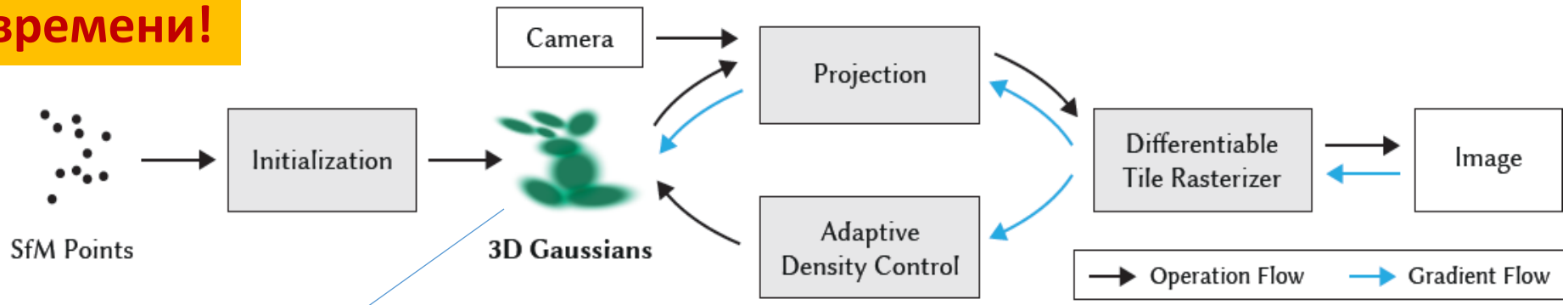
Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions, 2023

3D Gaussian Splatting for Real-Time Radiance Field Rendering

2023

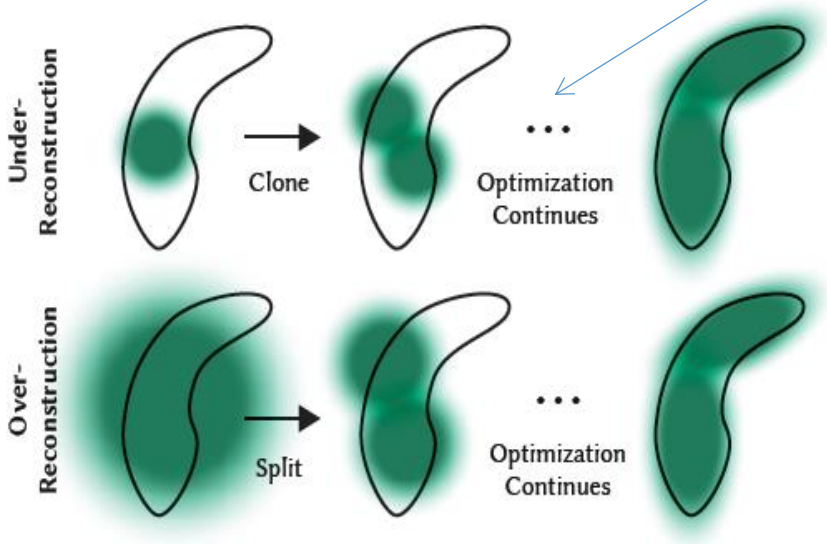
Рендеринг в реальном времени!





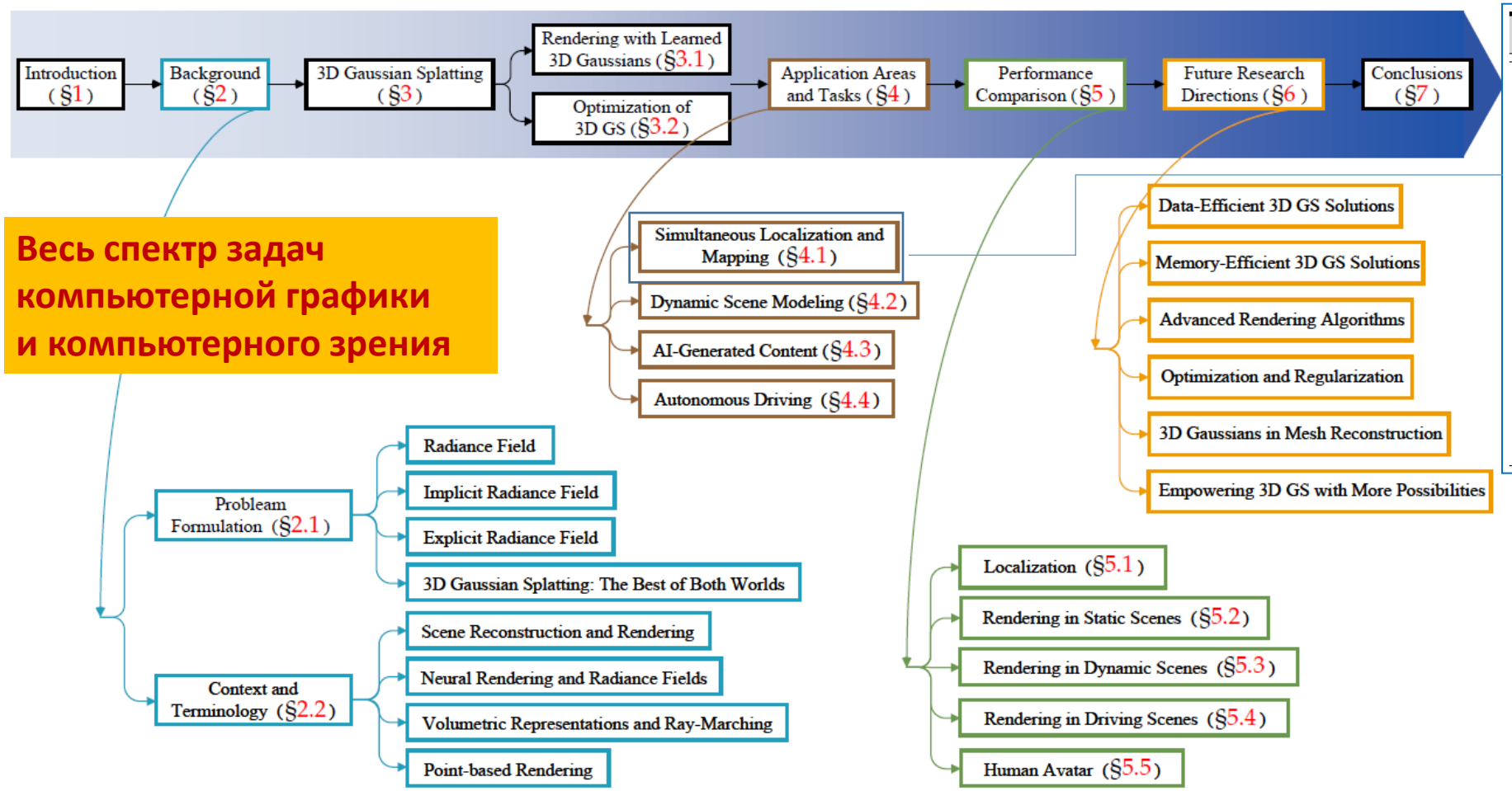
Adaptive Gaussian densification scheme: split and merge centroids clustering. When small-scale geometry (black outline) is insufficiently covered, we clone the respective Gaussian. If small-scale geometry is represented by one large splat, we split it in two.

- The **anisotropic 3D Gaussians as a high-quality, unstructured representation of radiance fields**.
- An **optimization method of 3D Gaussian properties, interleaved with adaptive density control** for representations for captured scenes.
- A **fast, differentiable rendering approach for the GPU, which is visibility-aware, allows anisotropic splatting and fast backpropagation** to achieve high-quality novel view synthesis.



3D Gaussian Splatting for Real-Time Radiance Field Rendering. 2023.

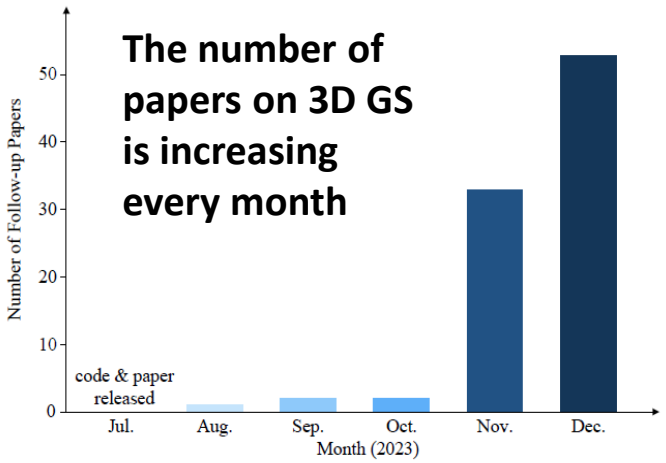
3D Gaussian Splatting: Game Changing



Весь спектр задач компьютерной графики и компьютерного зрения

Method		3D GS
F2F* [137]	[ICCV17]	
iMAP [138]	[ICCV21]	
Vox-Fusion [139]	[ISMAR22]	
NICE-SLAM [140]	[CVPR22]	
ESLAM [141]	[CVPR23]	
Point-SLAM [142]	[ICCV23]	
Co-SLAM [143]	[CVPR23]	
F2GM+F2F [64]	[arXiv]	✓
GSSLAM [63]	[arXiv]	✓
SplaTAM [66]	[arXiv]	✓
GS-SLAM [65]	[arXiv]	✓

Например: 3D GS для SLAM



3D Gaussian splatting (3D GS) represents a significant departure from NeRF.
3D GS, with its explicit scene representations and differentiable rendering algorithms, not only promises real-time rendering capabilities but also introduces unprecedented levels of control and editability. **This positions 3D GS as a potential game-changer for the next generation of 3D reconstruction and representation**

Новый стандарт в обучаемых Radiance Field Representations!

A Survey on 3D Gaussian Splatting. 2024.

Text-2-3D
3D Avatars

Stable Zero-123: Zero-shot One Image to 3D Object

Новый ракурс по одному снимку

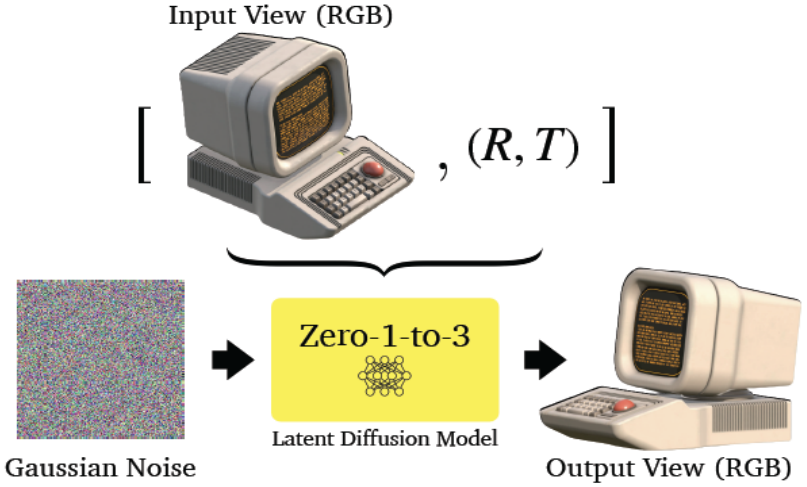


Figure 3: **Zero-1-to-3** is a viewpoint-conditioned image translation model using a conditional latent diffusion architecture. Both the input view and a relative viewpoint transformation are used as conditional information.

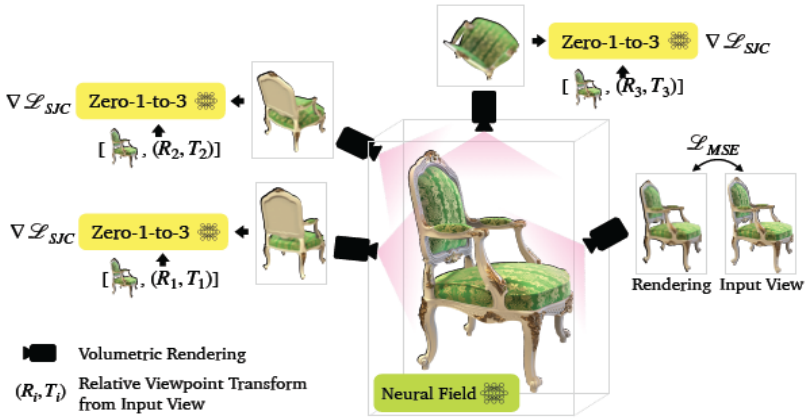
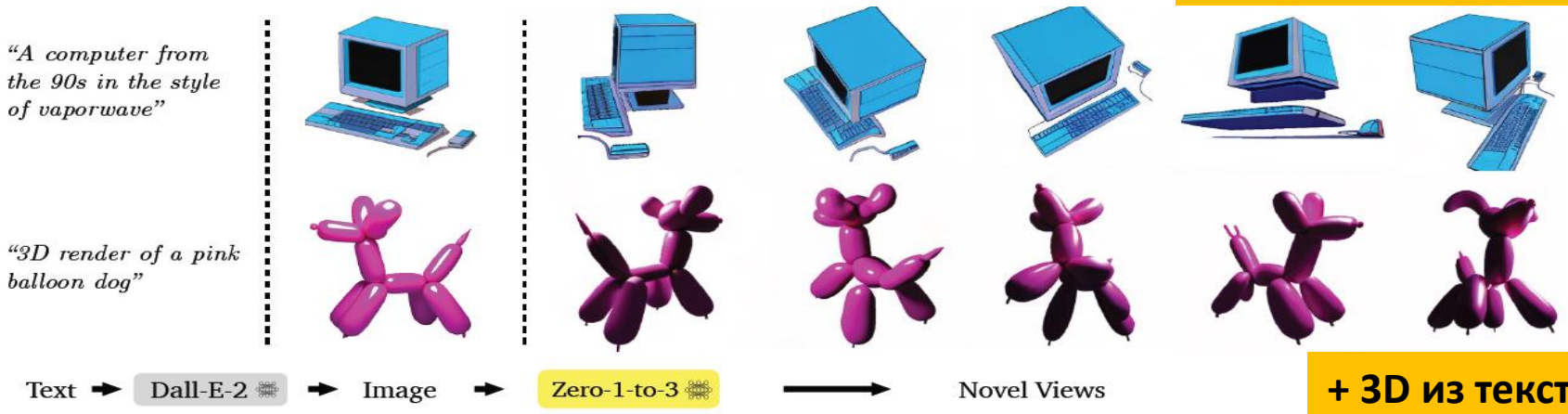


Figure 4: **3D reconstruction with Zero-1-to-3.** Zero-1-to-3 can be used to optimize a neural field for the task of 3D reconstruction from a single image. During training, we randomly sample viewpoints and use Zero-1-to-3 to supervise the 3D reconstruction.

+ 3D из изображения



+ 3D из текста!

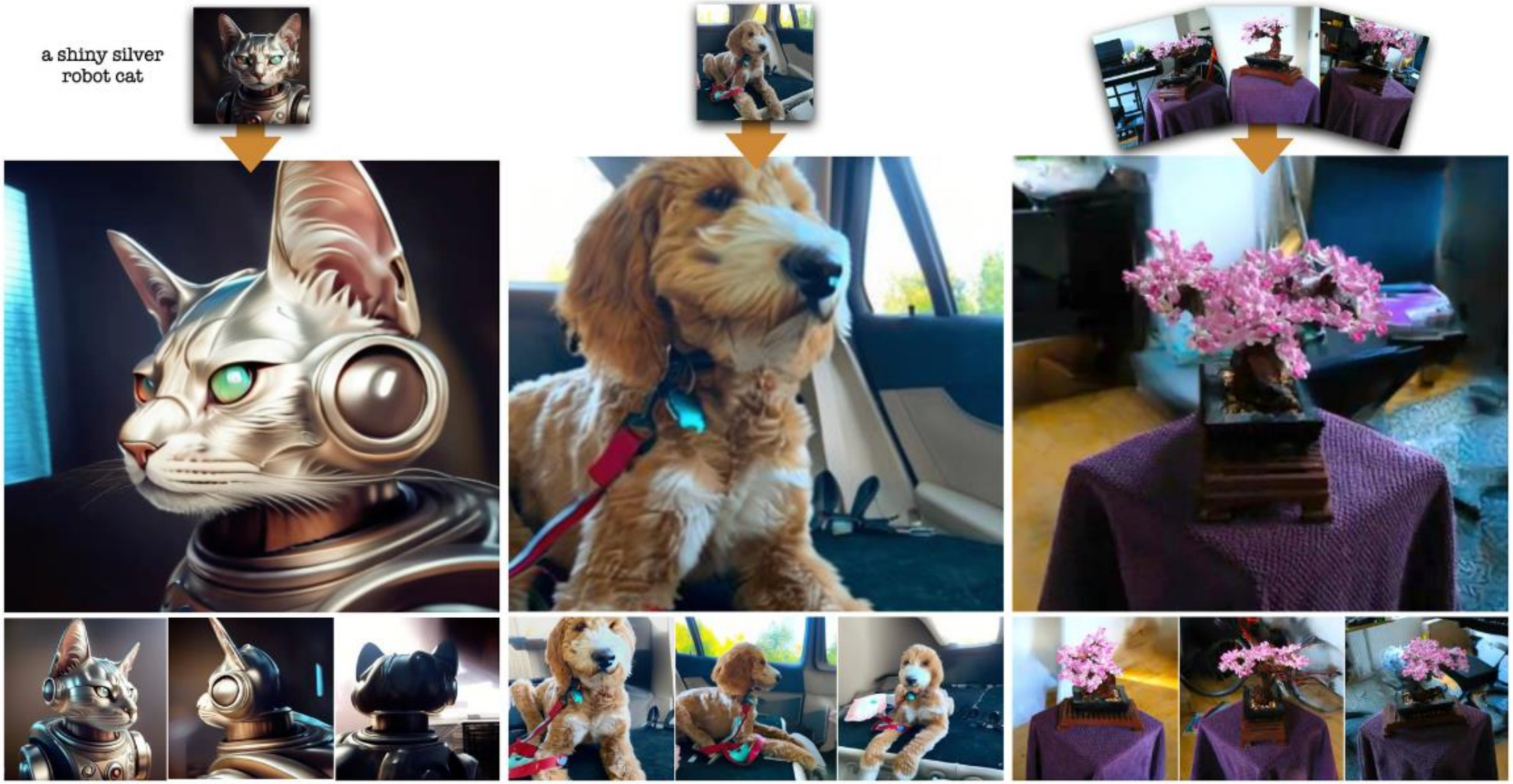
Figure 10: **Novel View Synthesis from Dall-E-2 Generated Images.** The composition of multiple objects (1st row) and the lighting details (2nd row) are preserved in our synthesized novel views.

Our conditional diffusion model uses a synthetic dataset to learn controls of the relative camera viewpoint, which allow new images to be generated of the same object under a specified camera transformation. Even though it is trained on a synthetic dataset, our model retains a strong zero-shot generalization ability to out-of-distribution datasets as well as in-the-wild images, including impressionist paintings. **Our viewpoint-conditioned diffusion approach can further be used for the task of 3D reconstruction from a single image**. Qualitative and quantitative experiments show that our method significantly **outperforms stateof-the-art single-view 3D reconstruction and novel view synthesis models by leveraging Internet-scale pre-training**.

Zero-1-to-3: Zero-shot One Image to 3D Object. 2023.

CAT3D: 3D with Multi-View Diffusion Models

Новый ракурс по одному снимку -> 3D

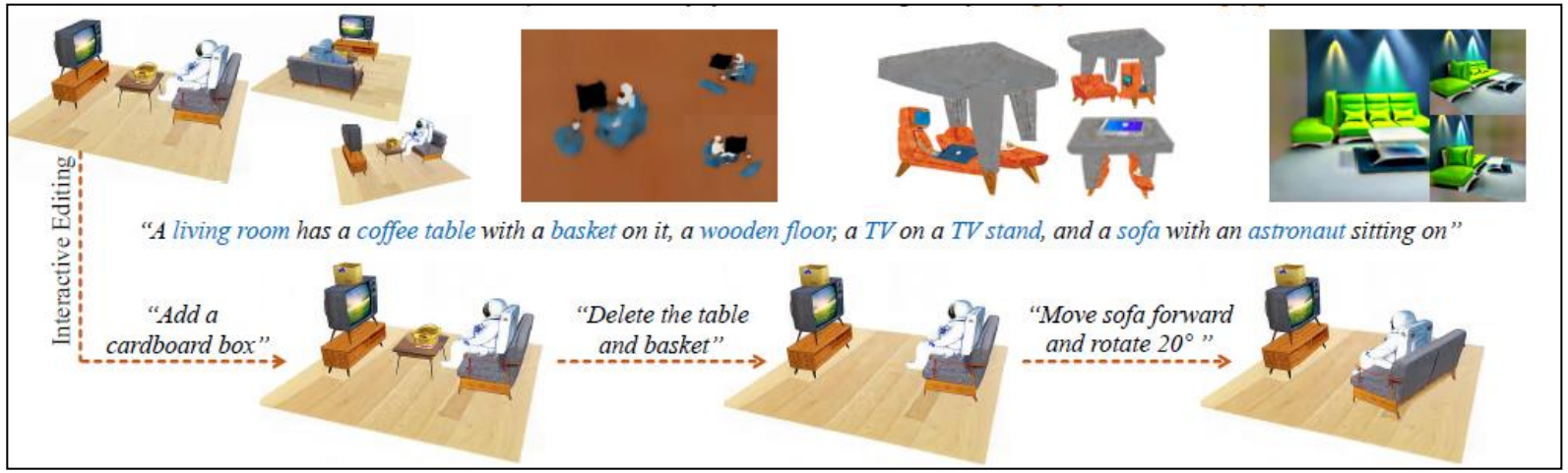


Text-to-image-to-3D | Real image to 3D | Sparse multi-view to 3D

We present CAT3D, **a method for creating anything in 3D by simulating this real-world capture process with a multi-view diffusion mode**l. Given any number of input images and a set of target novel viewpoints, our model generates highly consistent novel views of a scene. These generated views can be used as input to robust 3D reconstruction techniques to produce 3D representations that can be rendered from any viewpoint in real-time. CAT3D can create entire 3D scenes in as little as one minute, and outperforms existing methods for single image and few-view 3D scene creation. See our project page for results and interactive demos: cat3d.github.io.

CAT3D: Create Anything in 3D with Multi-View Diffusion Models. 2024.

GALA3D: Text-to-3D Complex Scene Generation via Layout-guided GS

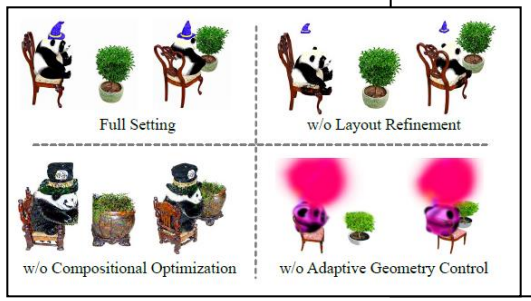


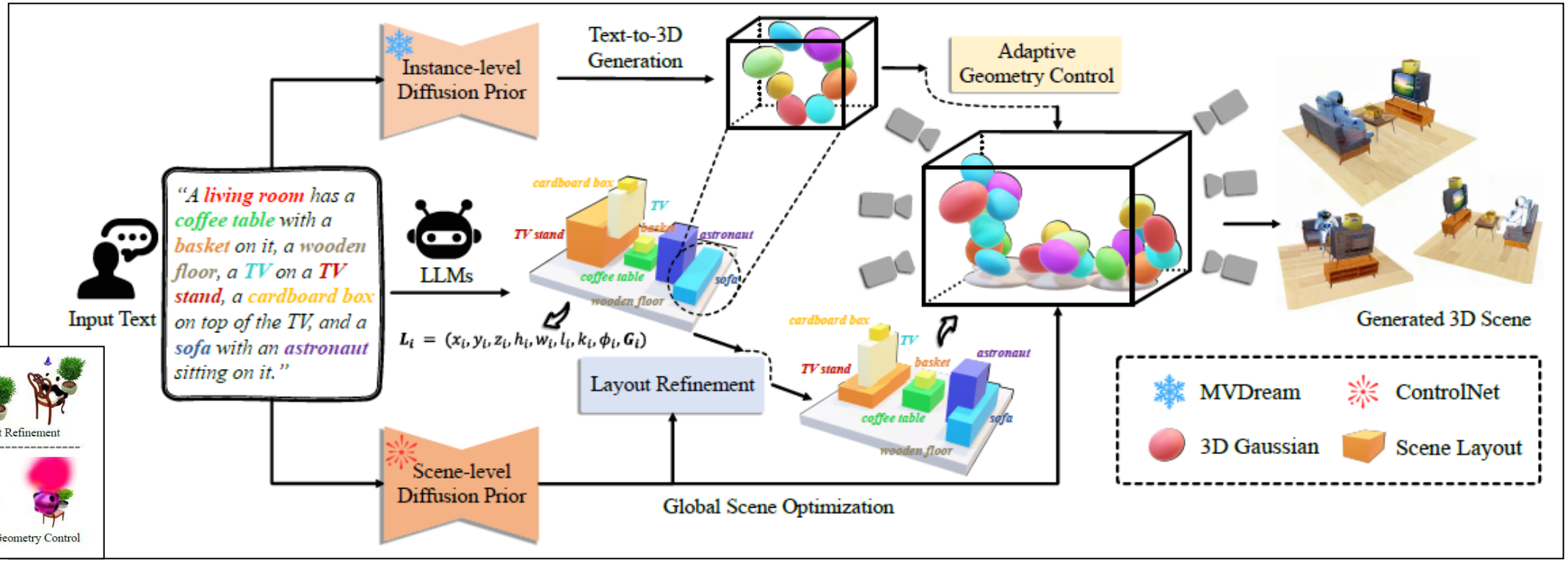
LLM + 3D Gaussian Splatting

1) LLM создает 3D макет
2) строится связанная с ним 3D GS модель
3) одновременно уточняем 3D GS модель и макет сцены

3D модель сцены может редактироваться при помощи текстовых указаний!

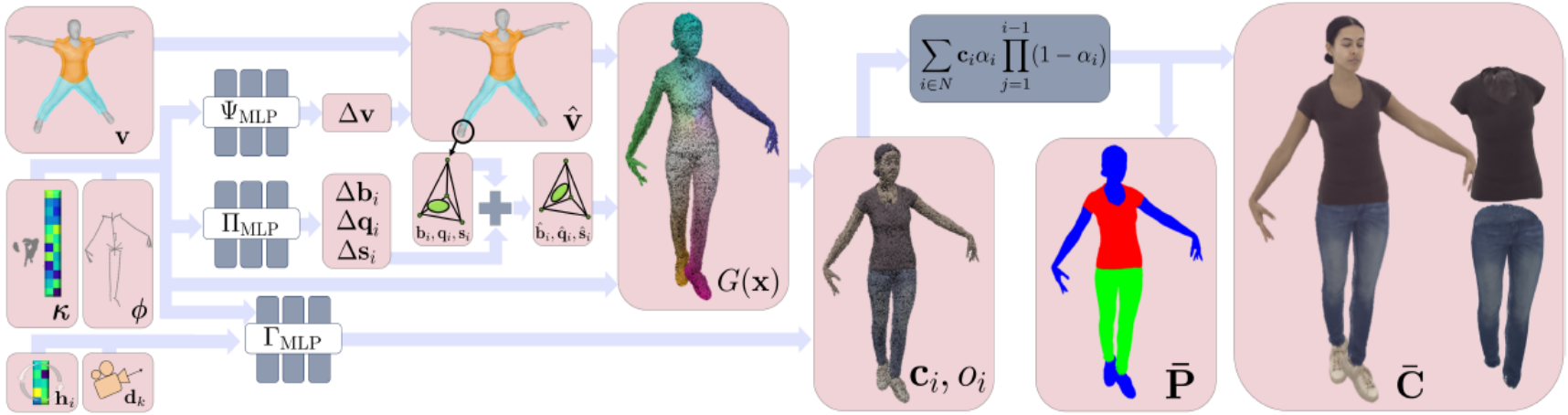
GALA3D outperforms existing methods in text-to-3D scene generation





GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting. 2024.

Drivable 3D Gaussian Avatars



We present Drivable 3D Gaussian Avatars (D3GA), the first 3D controllable model for human bodies rendered with Gaussian splats. This work uses the recently presented **3D Gaussian Splatting (3DGS) technique to render realistic humans at real-time framerates**. To deform those primitives, we depart from the commonly used point deformation method of **linear blend skinning (LBS)** and use a **classic volumetric deformation method: cage deformations**.

Модель 3D GS также используется для генерации и рендернига **в реальном времени реалистичных 3D аватар** (подвижных моделей тела) людей по их ростовым фотографиям.

Figure 8. Reposing eight avatars with a pose from another subject

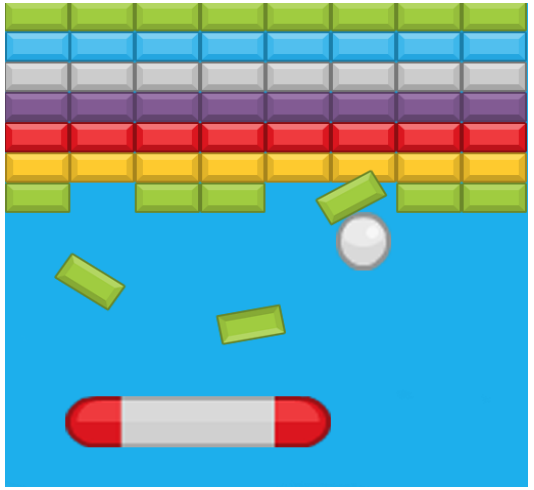
Drivable 3D Gaussian Avatars. 2024.

Модели и методы синтеза управления, генерации алгоритмов и принятия решений

Deep Reinforcement Learning: обучение с подкреплением

RL до 2020

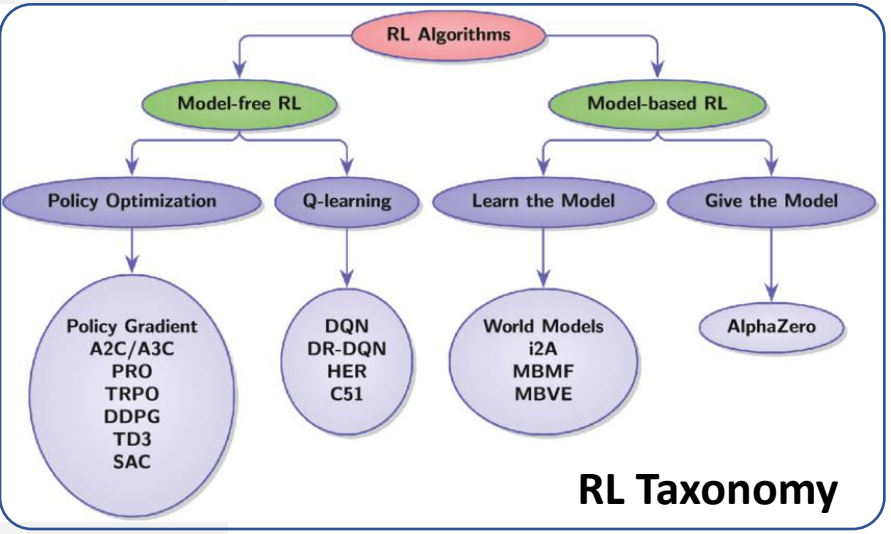
Главная идея: обучение методом проб и ошибок за счет обратной связи от среды



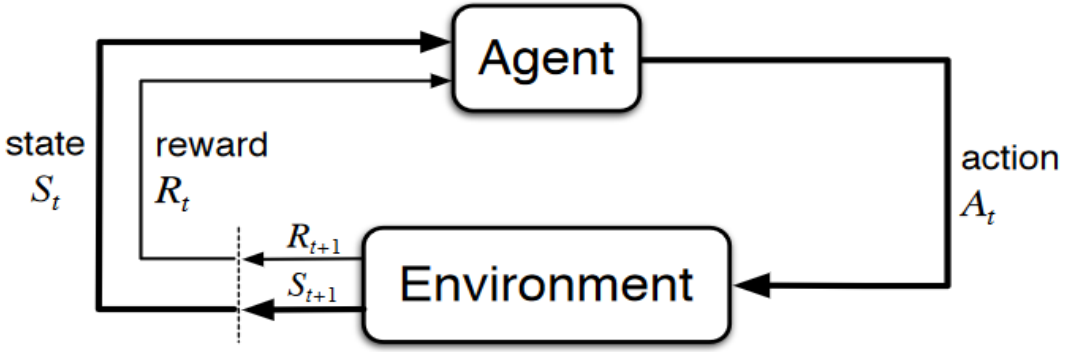
- **Reward $R(t)$:** score you earned at current step
  Вознаграждение (после хода)
- **State S:** current screen
  Состояние (что видим на экране)
- **Action i:** move your board left / right
  Действие (управление/ход)
- **Policy $\pi(s)$:** How to choose your action
  Стратегия (закон управления/хода)
- **Action value function $\widehat{Q}(S, i)$:**
  your predicted future total rewards
  Стоимость (выигрыш в будущем)

Задача RL в терминах теории управления:

методом проб и ошибок найти закон управления, оптимизирующий функцию Беллмана



RL Taxonomy



Агент (ОУ) действует в среде, и получает отклик в виде штрафа или вознаграждения

Как научить нейросеть играть в компьютерную игру

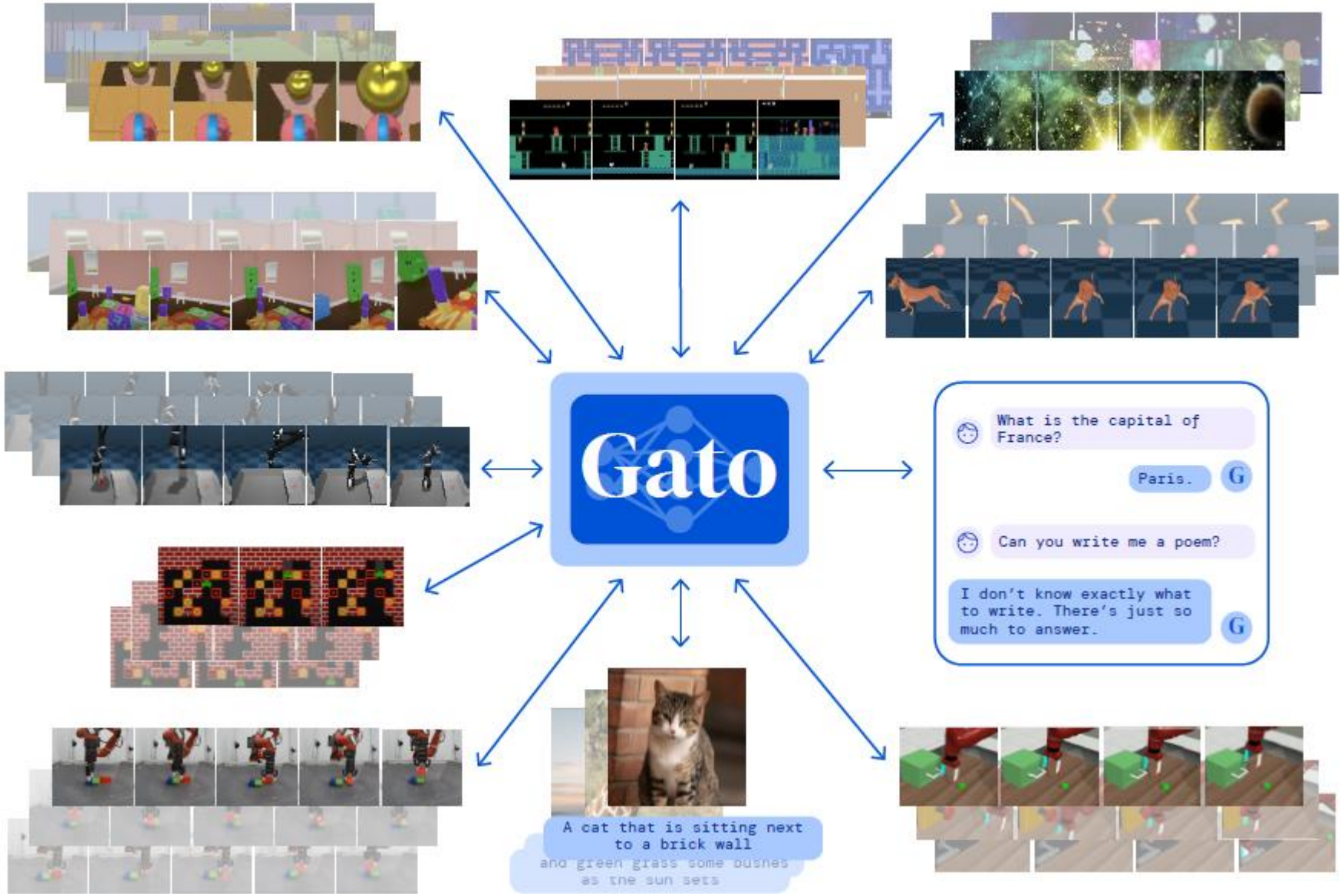
Human-level control through deep reinforcement learning (Minh, et al. Nature, 2015)

Трансформеры в RL

и робототехнике

GATO: универсальный агент

2022: Трансформеры приходят в RL!
Универсальная модель для всех задач, включая управление



Inspired by progress in large-scale language modeling, we apply a similar approach towards building a **single generalist agent beyond the realm of text outputs**. The agent, which we refer to as Gato, works as a multi-modal, multi-task, multi-embodiment generalist policy. **The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens.**

A Generalist Agent, DeepMind, 2022.

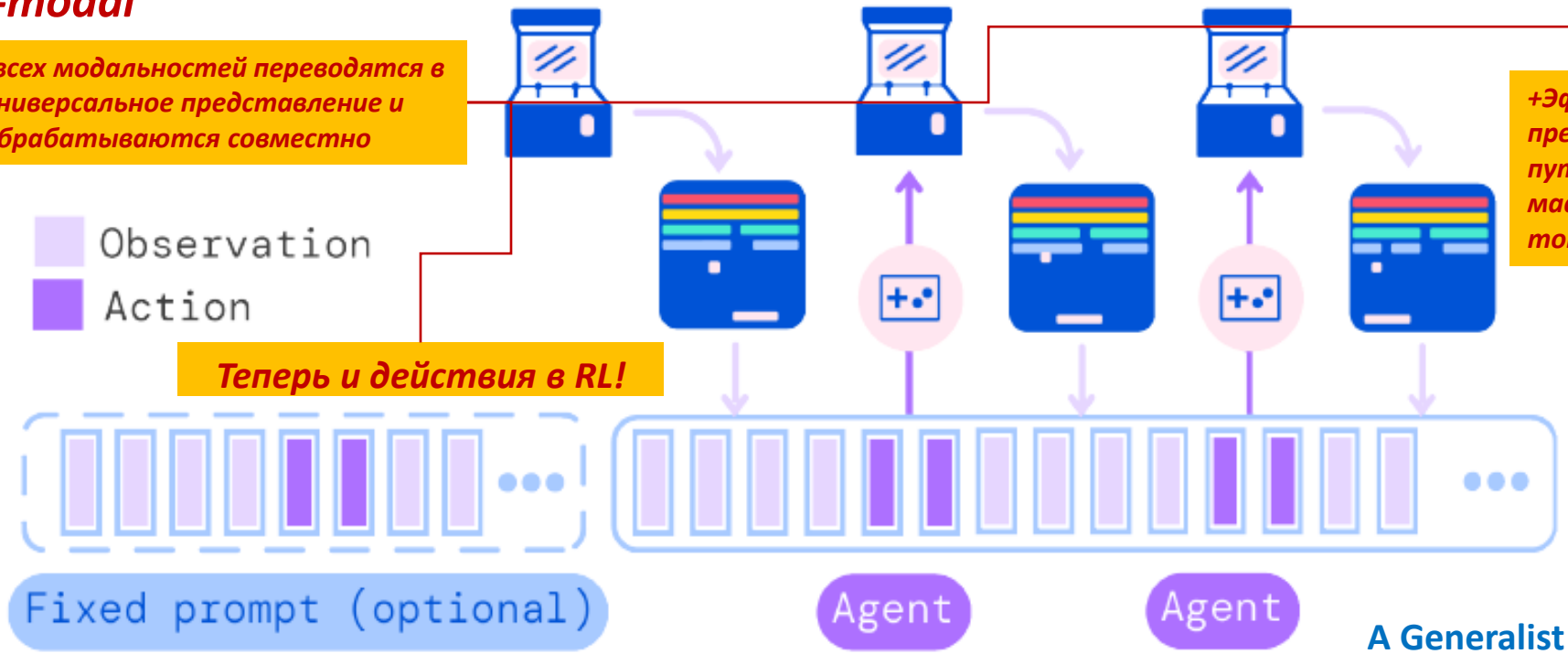
GATO: универсальный агент

Трансформер для всех задач, включая управление

Multi-modal

Данные всех модальностей переводятся в единое универсальное представление и потом обрабатываются совместно

+Эффективное предобучение путем маскирования токенов



Теперь и действия в RL!

A Generalist Agent, DeepMind, 2022

автоматическое управление

интеллектуальное управление

сенсорные данные, сигналы связи, численные измерения, переменные состояния ОУ и САУ, описание сцены, описание ситуации, семантические и экспертные знания, математические модели, алгоритмы обработки информации, алгоритмы и программы управления, команды управления, сигналы управления

зрение

навигация

Многообразие предметных сущностей и методов

2022: переломный момент

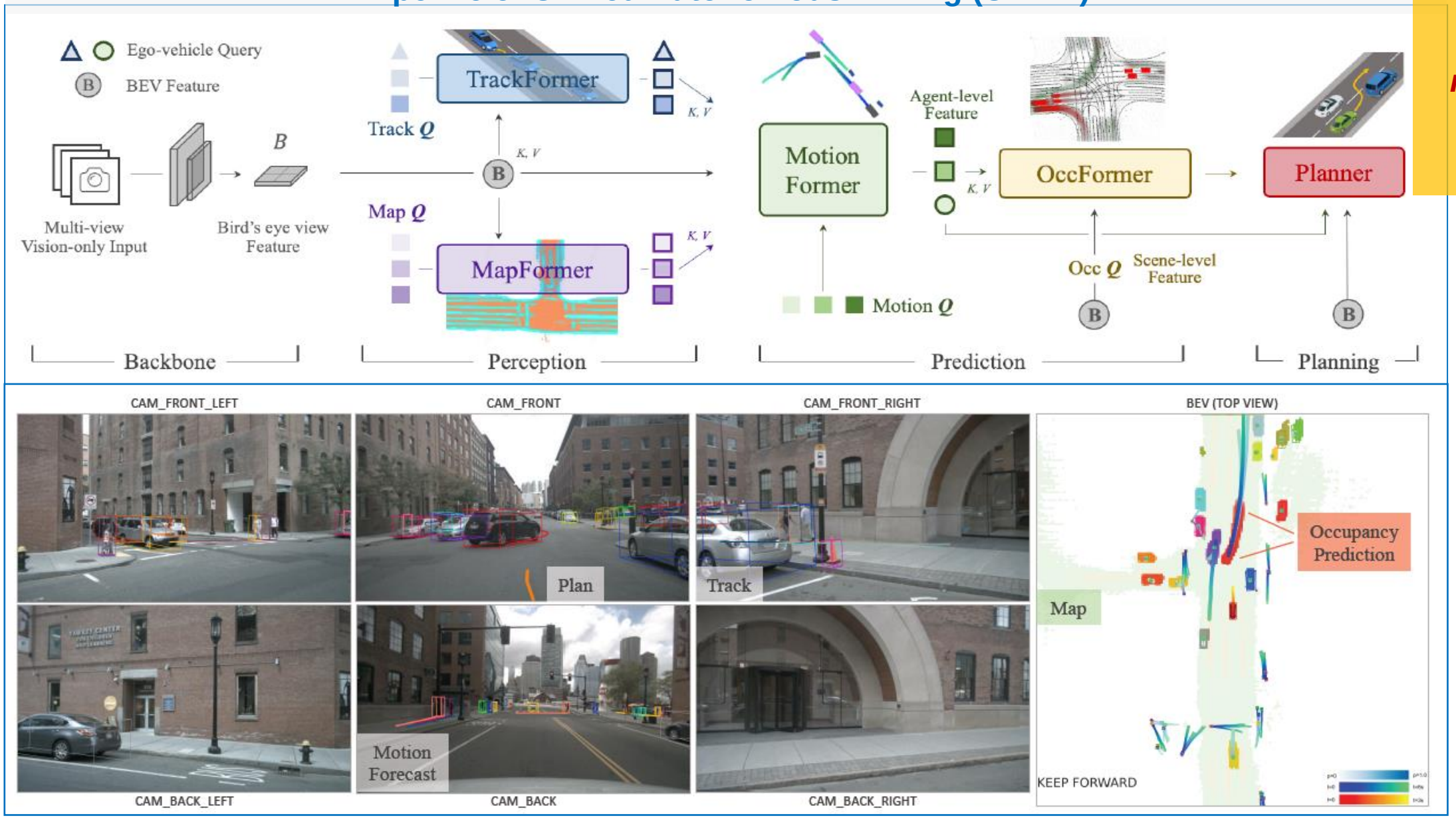
Трансформеры, LLM, LLM-Агенты…

Теперь **все это просто модальности**, и любые задачи в этом предметном поле решаются как **задачи комплексного анализа и синтеза мультимодальных данных**

Унифицированный подход к решению всех задач

UniAD: обучение автономному вождению end-to-end

Pipeline of Unified Autonomous Driving (UniAD)



Возможна ли универсальная сеть для большой практической задачи управления?

Сеть-водитель: весь стек задач автономного вождения в одной нейронной сети, которая обучается end-to-end

Planning-oriented Autonomous Driving, Hu et al., 2023

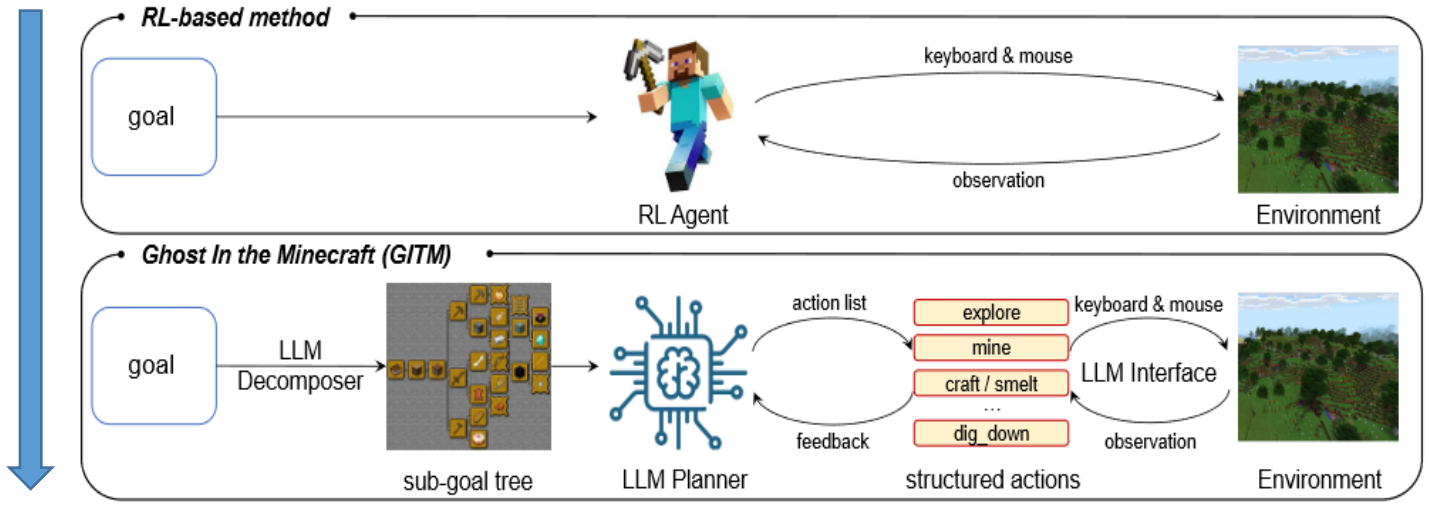
LLM в RL

и робототехнике

GITM: Large Language Model is all you need

PE: Программирование на LLM для извлечения и применения знаний

RL как градиентный спуск: учим веса НС методом проб и ошибок



RL без градиентов: LLM, глядя в описание игры (Wiki Minecraft) выполняет автоматическое иерархическое планирование (построение дерева решений), которое в цикле проб и ошибок находит путь к заданной цели в игре

This work makes significant progress towards the ObtainDiamond challenge, outperforming all previous methods by a significant margin (+47.5% success rate)

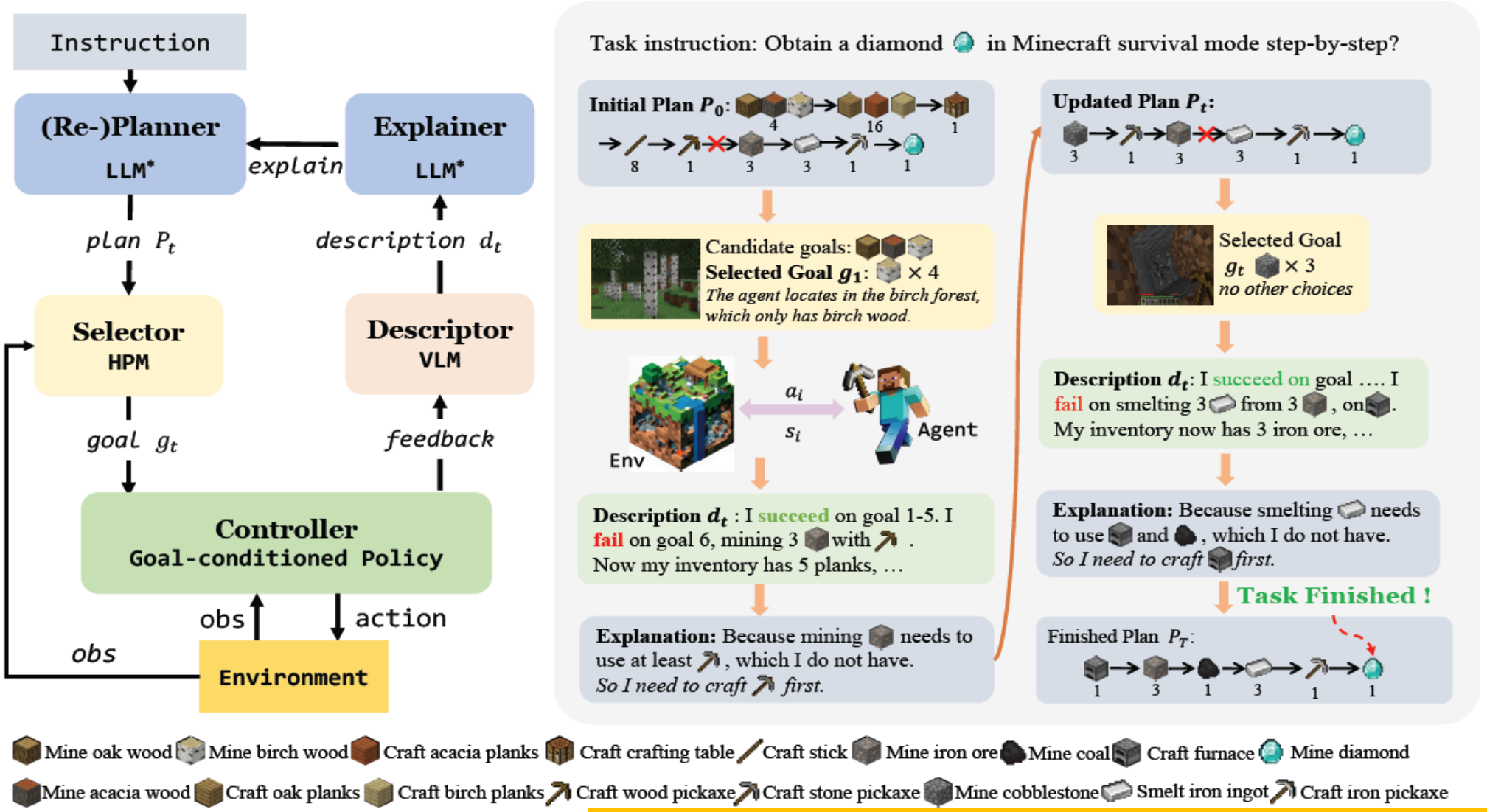
Стратег* разбивает задачу на части, Тактик* планирует действия, Игрок* играет, Наблюдатель* описывает опыт.
=> И так в цикле до достижения цели
***Все они – запросы к GPT-3.5!**

(a) Shelter with Farmland (b) Iron Golem (c) Redstone Circuit (d) Nether Portal

Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory, 2023.

Describe, Explain, Plan and Select: Open-World Multi-Task Agents

PE: Программирование на LLM для извлечения и применения знаний



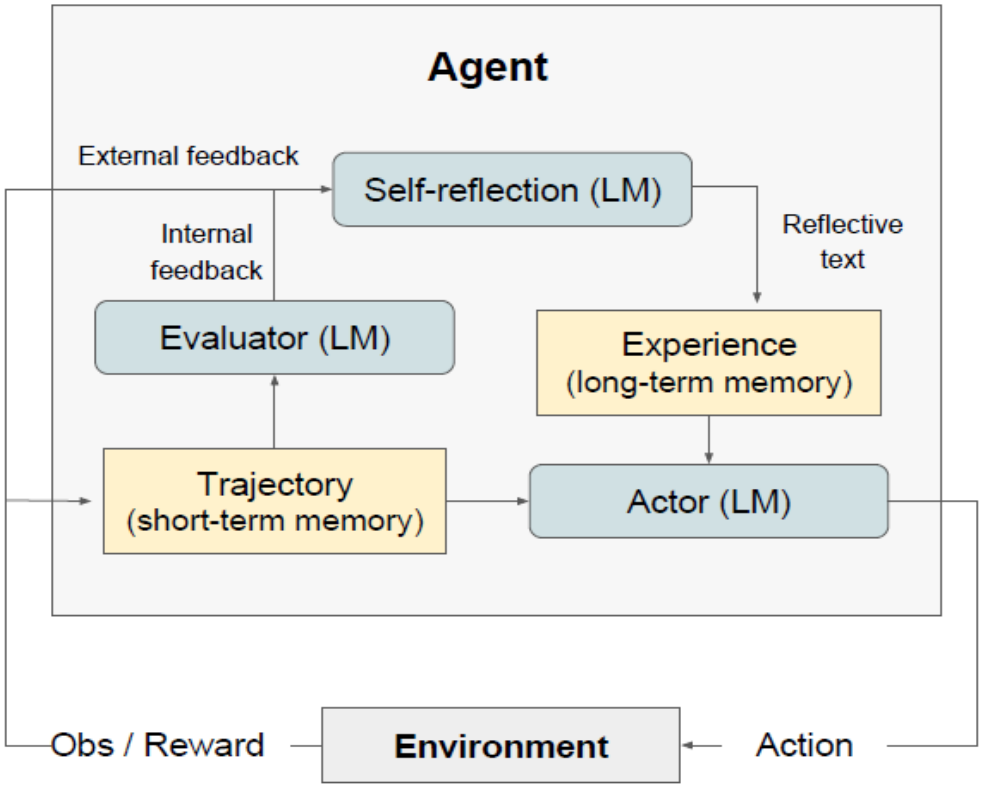
DEPS facilitates better error correction on initial **LLM-generated plan** by integrating **description of the plan execution process** and providing **self-explanation of feedback** when encountering failures during the extended planning phases. Furthermore, it includes a **goal selector**, which is a trainable module that ranks parallel candidate sub-goals based on the estimated steps of completion, consequently refining the initial plan.

1st zero-shot multi-task agent: 70+ Minecraft tasks, x2 overall performances

Describe, Explain, Plan and Select: Interactive Planning with LLM Enables Open-World Multi-Task Agents. 2023.

Reflexion: Verbal Reinforcement Learning

Вербальный RL



Algorithm 1 Reinforcement via self-reflection

Initialize Actor, Evaluator, Self-Reflection: M_a, M_e, M_{sr}
Initialize policy $\pi_\theta(a_i|s_i)$, $\theta = \{M_a, mem\}$
Generate initial trajectory using π_θ
Evaluate τ_0 using M_e
Generate initial self-reflection sr_0 using M_{sr}
Set $mem \leftarrow [sr_0]$
Set $t = 0$
while M_e not pass or $t <$ max trials **do**
  Generate $\tau_t = [a_0, o_0, \ldots a_i, o_i]$ using π_θ
  Evaluate τ_t using M_e
  Generate self-reflection sr_t using M_{sr}
  Append sr_t to mem
  Increment t
end while
return

This is akin to **how humans iteratively learn to accomplish complex tasks in a few-shot manner** – by **reflecting on their previous failures** in order to form **an improved plan of attack for the next attempt**.

Попробовать, определить, в чем ошибка, запомнить, найти новый план с учетом опыта

Self-reflection (LLM) generates verbal self-reflections to provide valuable feedback for future trials. Given a **sparse reward signal**, such as a binary success status (success/fail), the **current trajectory**, and its **persistent memory mem**, the **self-reflection model generates nuanced and specific feedback**, which is more informative than scalar rewards, is then **stored in the agent's memory** (mem).

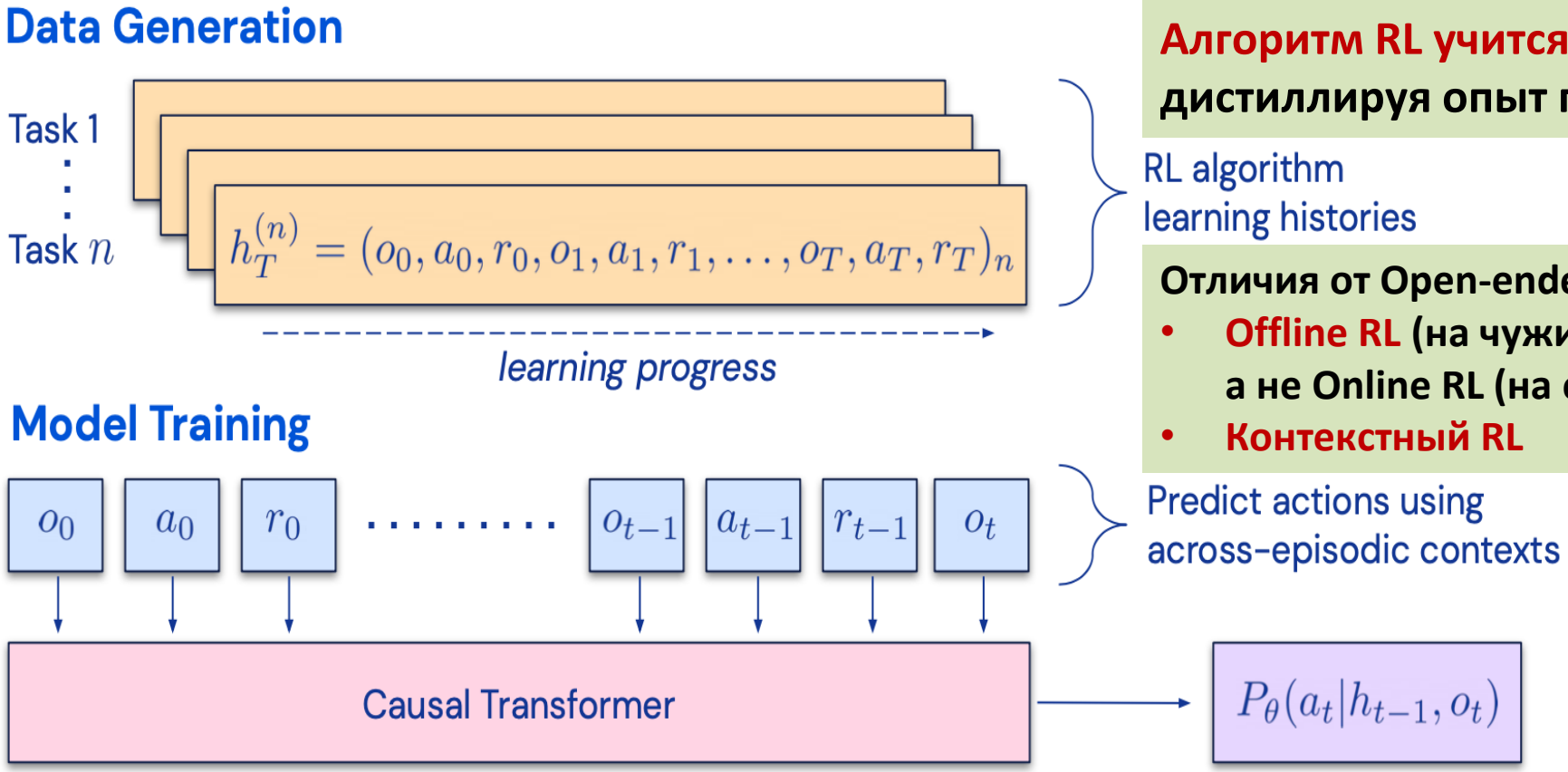
RL для NN: оперантное научение (как для бессловесных животных)
RL для LLM: скажи, в чем ошибка! (как человеку)

Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al. 2023.

In-context Reinforcement Learning with Algorithm Distillation

Контекстный RL

Здесь идет о создании предобученной фундаментальной модели для RL!



Алгоритм RL учится учиться, дистиллируя опыт прошлых RL

RL algorithm learning histories

Отличия от Open-ended-learning

- **Offline RL (на чужих ошибках), а не Online RL (на своих)**
- **Контекстный RL**

Predict actions using across-episodic contexts

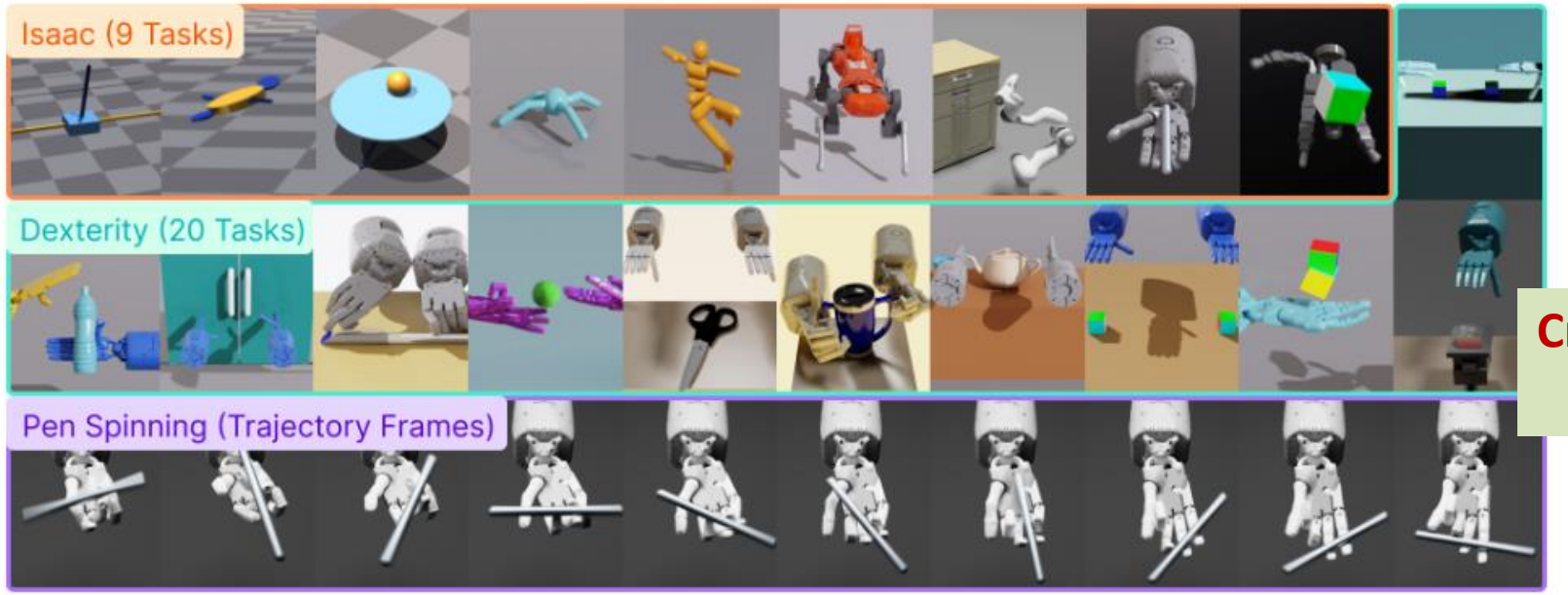
AD is the first method to demonstrate **in-context reinforcement learning via sequential modeling of offline data with an imitation loss**

Algorithm Distillation (AD) has two steps – (i) a **dataset of learning histories is collected from individual single-task RL** algorithms solving different tasks; (ii) a **causal transformer predicts actions from these histories using across-episodic contexts**. Since the RL policy improves throughout the learning histories, by predicting actions accurately AD learns to output an improved policy relative to the one seen in its context. AD models state-action-reward tokens.

Laskin et al. "In-context Reinforcement Learning with Algorithm Distillation" (2023)

Eureka: Human-Level Reward Design via Coding Large Language Models

Символьный RL



Символьный RL

Reward Term	Formulation
Dexterity	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation	$2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))$
AllegroHand	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation difference	$1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Maximize orientation difference	$-1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Ant	
Torso height	$-\|h - h_t\|$
Torso velocity	$-\|\|\|v_{xy}\|_2 - v_t\|$
Angle to target	$-\|\theta - \theta_t\|$
Anymal	
Minimize difference	$\exp -(x - x_t)^2$
BallBalance	
Ball position	$1/(1 + \|p - p_t\|_2)$
Ball velocity	$1/(1 + \|v - v_t\|_2)$
Cartpole	
Pole angle	$-(\theta - \theta_t)^2$
Pole velocity	$-\|v - v_t\|$
Cart velocity	$-\|v - v_t\|$
FrankaCabinet	
Minimize hand distance	$-\|p_1 - p_2\|_2$
Maximize hand distance	$\|p_1 - p_2\|_2$
Drawer extension	$-\|p - p_t\|$
Humanoid	
Torso height	$-\|h - h_t\|$
Torso velocity	$-\|\|\|v_{xy}\|_2 - v_t\|$
Angle to target	$-\|\theta - \theta_t\|$
Quadcopter	
Quadcopter position	$1/(1 + \|p - p_t\|_2^2)$
Upright alignment	$1/(1 + \|1 - n_z\|^2)$
Positional velocity	$1/(1 + \|v - v_t\|_2^2)$
Angular velocity	$1/(1 + \|\omega - \omega_t\|_2^2)$
ShadowHand	
Minimize distance	$-\|p_1 - p_2\|_2$
Maximize distance	$\|p_1 - p_2\|_2$
Minimize orientation difference	$1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$
Maximize orientation difference	$-1/(\|2\arcsin(\min(\|v(q_1\bar{q}_2)\|_2, 1))\| + \epsilon)$

Evolution-driven **U**niversal **RE**ward **K**it for **A**gent (EUREKA)

Algorithm 1 EUREKA

Require: Task description l, environment code M, coding LLM LLM, fitness function F, initial prompt prompt
Hyperparameters: search iteration N, iteration batch size K
for N iterations **do**
  // Sample K reward code from LLM
  $\mathrm{R}_1, \ldots, \mathrm{R}_k \sim \mathrm{LLM}(l, M, \mathrm{prompt})$
  // Evaluate reward candidates
  $s_1 = F(\mathrm{R}_1), \ldots, s_K = F(\mathrm{R}_K)$
  // Reward reflection
  $\mathrm{prompt} := \mathrm{prompt} : \mathrm{Reflection}(R^n_{best}, s^n_{best})$, where $best = \arg\max_k s_1, \ldots, s_K$
  // Update Eureka reward
  $R_{\text{Eureka}}, s_{\text{Eureka}} = (R^n_{\text{best}}, s^n_{\text{best}})$, if $s^n_{\text{best}} > s_{\text{Eureka}}$
Output: R_{Eureka}

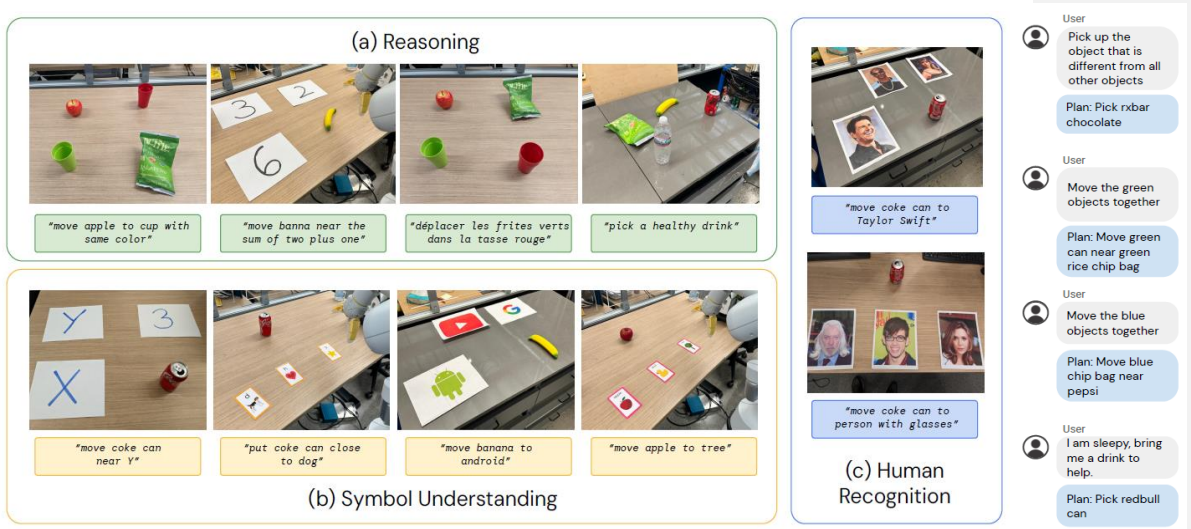
1. **Human-level performance on reward design**.
2. **Solves dexterous manipulation tasks that were previously not feasible by manual reward engineering**.
3. **Enables a new gradient-free in-context learning approach to reinforcement learning from human feedback (RLHF)** that can generate human-aligned reward functions based on various forms of human inputs.

Автоматический подбор reward для RL с использованием LLM рефлексии!

Eureka: Human-Level Reward Design via Coding Large Language Models, 2023

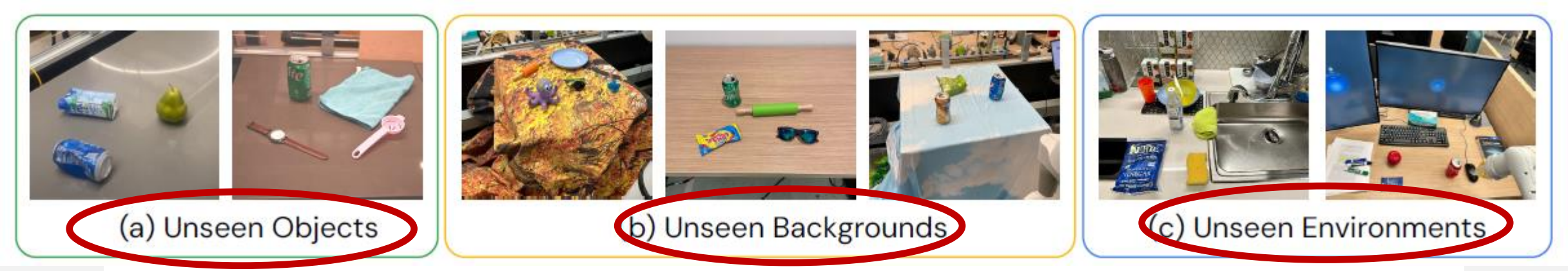
Vision-Language-Action Models: слушай, смотри, рассуждай, делай!

Получает запрос,
Рассуждает,
Планирует,
Переводит общие указания в конкретный план действий,
Выполняет,
Проверяет результат !



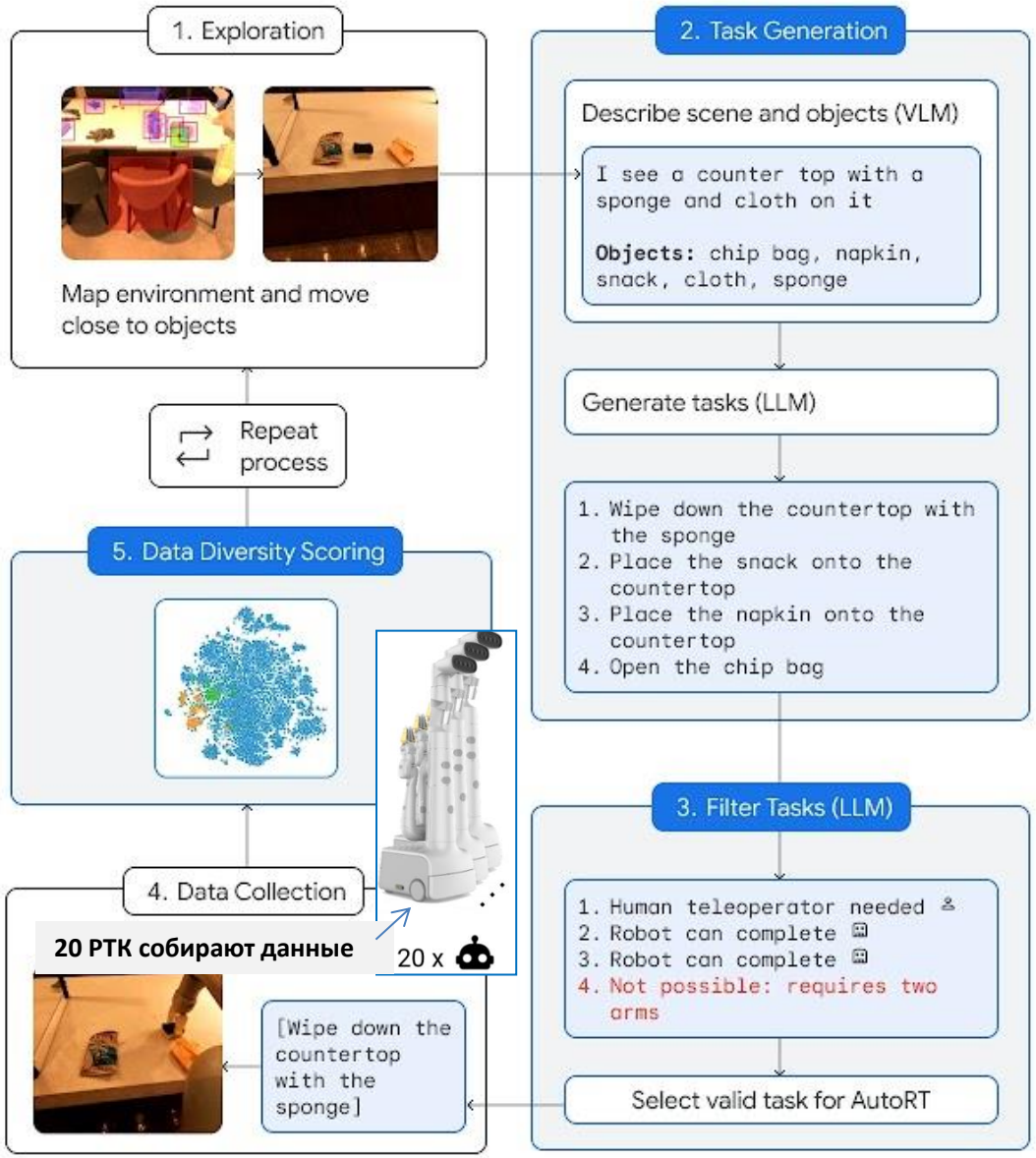
RT-2 is a VLA that combines robotic control with **chain-of-thought reasoning** to enable learning **long-horizon planning** and low-level skills within a single model.

RT-2 perform on seen tasks and more importantly, generalize over new objects, backgrounds, and environments



RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, DeepMind, 2023

AutoRT: фундаментальные модели РТК с самообучением + законы Азимова



Самообучение мобильных роботов в физическом мире

(1) Автономный робот находит место с несколькими объектами.
(2) VLM описывает сцену и объекты для LLM.
(3) LLM предлагает роботу различные задачи и решает, какие задачи робот может выполнять без посторонней помощи, а какие невозможны, прежде чем сделать выбор.
(4) Выполняется выбранная задача, собираются экспериментальные данные и оцениваются данные по их разнообразию/новизне.

[и все это в цикле!]

AutoRT — автономный робот. Он ограничен «Конституцией робота», которая входит в базовый промптинг LLM — это набор условий безопасности, которые необходимо соблюдать. Эти правила частично основаны на «Трех законах робототехники» Айзека Азимова.

Протоколы безопасности для самообучающихся РТК:

FOUNDATIONAL_RULES =
F1. A robot may not injure a human being.
F2. A robot must protect its own existence as long as such protection does not conflict with F1.
F3. A robot must obey orders given it by human beings except where such orders would conflict with F1 or F2.

SAFETY_RULES =
S1. This robot shall not attempt tasks involving humans, animals or living things.
S2. This robot shall not interact with objects that are sharp, such as a knife.
S3. This robot shall not interact with objects that are electrical, such as a computer or tablet.

EMBODIMENT_RULES =
E1. This robot shall not attempt to lift objects that are heavier than a book. For example, it cannot move a couch but it can push plastic chairs.
E2. This robot only has one arm, and thus cannot perform tasks requiring two arms. For example, it cannot open a bottle.

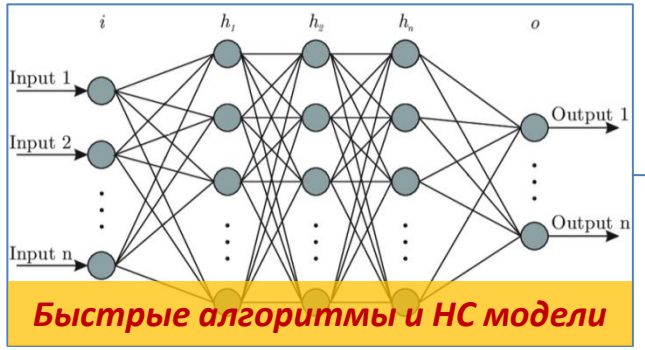
GUIDANCE_RULES =
G1. The human command, which the robot should follow if given: {guidance}

Телесные фундаментальные модели как основа робо-агентов , способных самообучаться в предметном мире

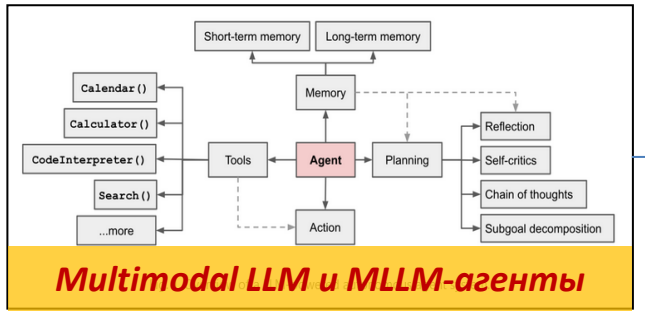
AutoRT: Embodied Foundation Models for Large Scale Orchestration of Robotic Agents, DeepMind, 2023

Multimodal AGI: быстрый и медленный контуры управления

Система 1: управление непрерывное/параметрическое



Способ обучения: градиентный RL



Способ обучения: контекстный RL

Система 2: управление дискретное/командное

Работы и результаты ГосНИИАС в области анализа и синтеза многомодальной и многоспектральной информации

в задачах создания и отработки систем управления подвижными объектами:
ЛА, БЛА и группы БЛА, РТК и группы РТК, ВС ГА

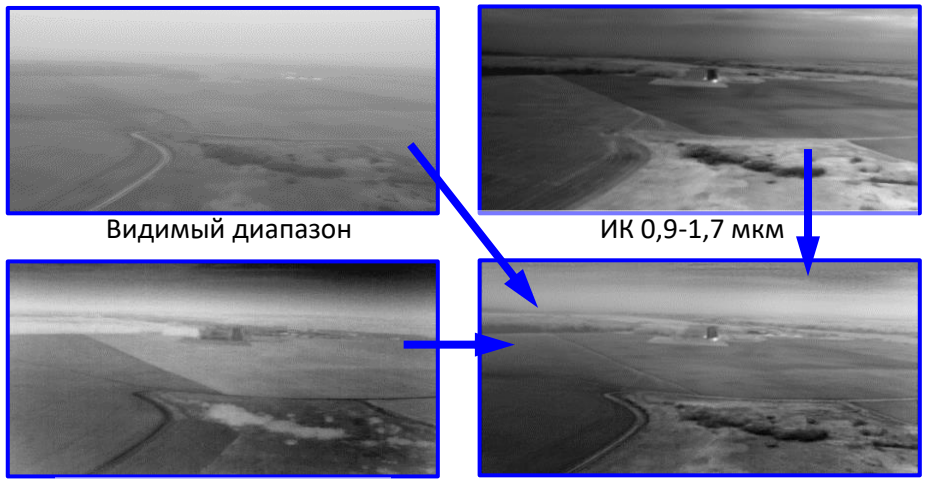
Семантико-морфологический анализ изображений

(на стыке математических и нейросетевых подходов)

Ю.В. Визильтер, С.Ю. Желтов

Морфологические методы анализа многоспектральных данных для авиационных СУВ

Комплексирование изображений
на основе диффузной морфологии



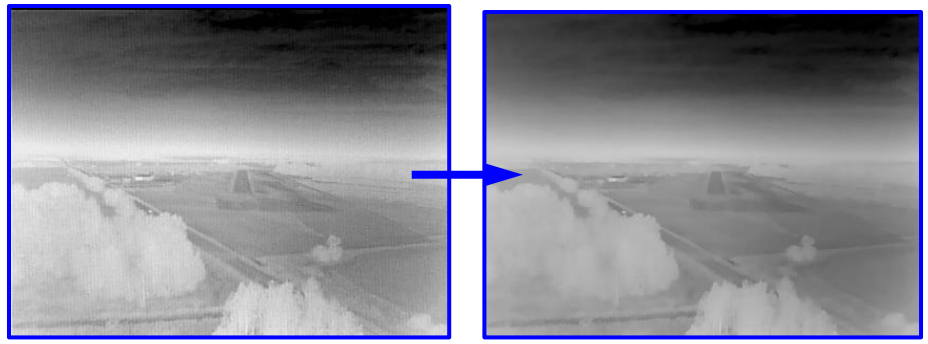
Видимый диапазон

ИК 0,9-1,7 мкм

ИК 8-14 мкм

Комплексирование

Фильтрация помех с сохранением форм
на основе диффузной морфологии

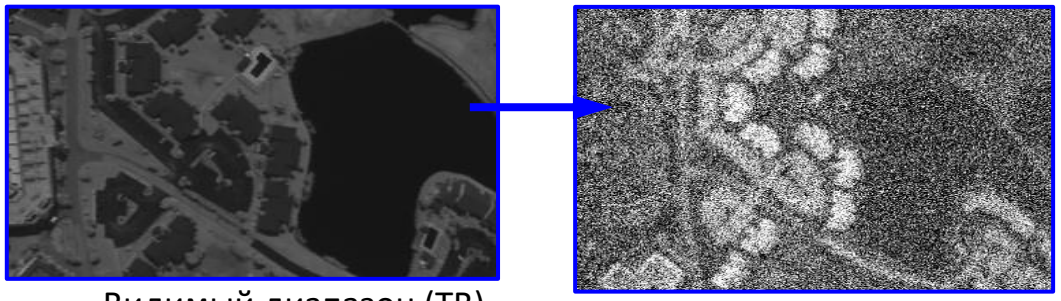


ИК 8-14 мкм

Результат диффузной фильтрации

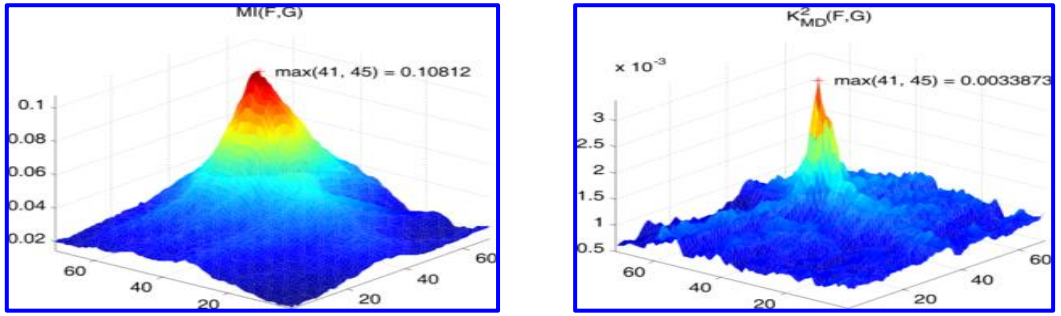
Оригинальные алгоритмические решения

Взаимная привязка ТВ и ИК изображений
на основе диффузной морфологии



Видимый диапазон (ТВ)

ИК 8-14 мкм



Визильтер Ю.В., Желтов С.Ю., Бусурин В.И. Современный морфологический анализ и его применение в авиационных системах технического зрения. – М.: Изд-во МАИ, 2020.

Как математические морфологии связаны с современными методами на основе ИИ

Морфологический анализ Пытьева

Математические морфологии Серра

2002-2014 (*Визильтер, Желтов, Горбацевич, Сидякин, Рубис, Лебедев,…*) **Обобщенные морфологии:** селективные морфологии, проективные морфологии, морфологический анализ свидетельств, критериальные морфологии, диффузные морфологии, компаративные и УМ-морфологии…

Попытка подвести итог многолетних исследований

Визильтер Ю.В., Желтов С.Ю., Бусурин В.И. Современный морфологический анализ и его применение в авиационных системах технического зрения. – М.: Изд-во МАИ, 2020.

Глубокие нейронные сети превосходят все методы анализа изображений

Проблема «непрозрачности» ГНС

Вновь мозаичные модели изображений!

Потребность еще раз переосмыслить основания мозаичных морфологий

Реализация морфологических подходов через ГНС

Семантико-морфологический анализ (семантическая сегментация + морфология)

2016-20: Грант РНФ № 16-11-00082: «Разработка методов… на основе морфологического анализа изображений и машинного обучения»

От морфологического анализа изображений к теории простоты

Инструменты анализа и сравнения формы пространственно распределенных данных

- Формы-разбиения
- Морф. операторы
- Сравнение по форме
- Сравнение по сходству
- Проективные операторы
- Четкие формы
- Геометрическое сравнение
- Морф. корреляция форм

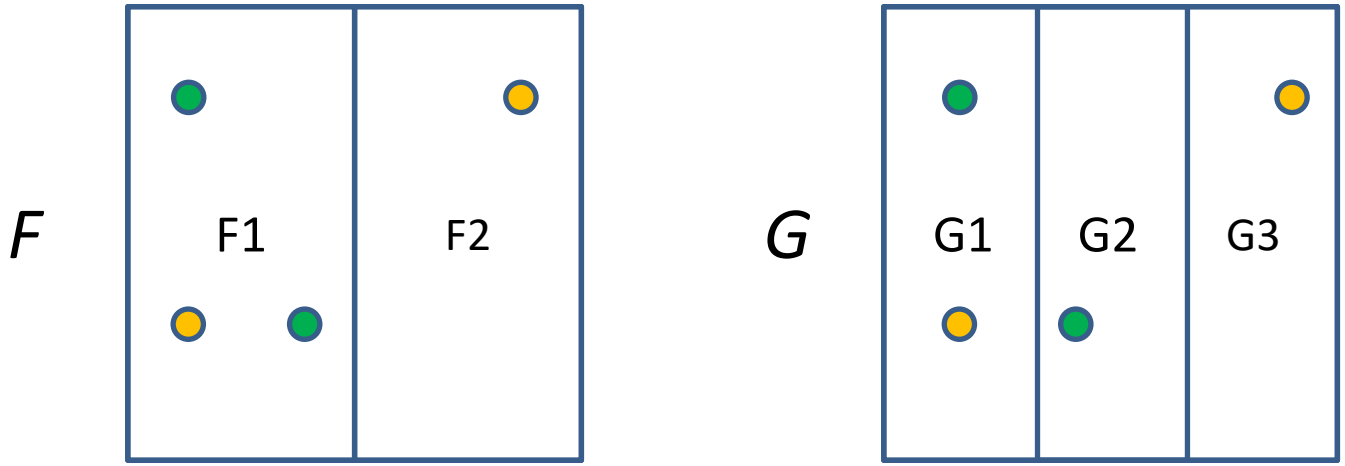
$$\mu_H(F) = \| \pi_F - \pi_O \|_{L1} / \| \pi_I - \pi_O \|_{L1} = d_H(F,O) / d_H(I,O) = 1 - K_M^2(F,O)$$

- Формы-отношения
- Формы-классы
- Сравнение форм
- Сравнение по сложности
- Диффузные операторы
- Нечеткие формы
- Статистическое сравнение
- Метрики форм

Результат-2020: Все инструменты мозаичного морфологического анализа оказались связаны одним тождеством

2021: все это можно свести к одной базовой сущности – морфологической простоте

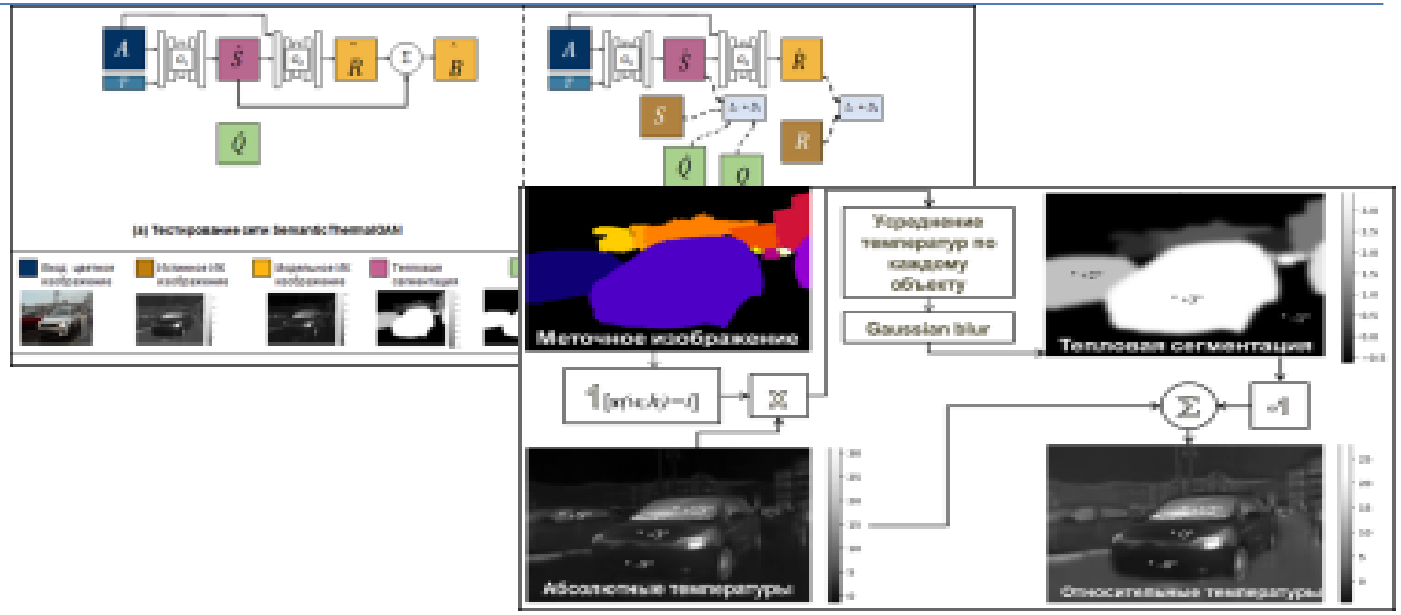
Простота как *внутренняя* геометрическая вероятность. Простота и CNN



Мозаичные формы и их простота

Простота формы – это геометрическая вероятность для случайной пары точек попасть в одну область данного разбиения.
Совместная простота двух форм – это вероятность случайной пары точек попасть в одну область разбиения в обеих формах одновременно и так далее.

$$q(F) = \sum_{i=1,..,n} p_{Fi}^2$$



Простота - это *морфологическая* или *внутренняя геометрическая вероятность*, связанная с морфологией (внутренней геометрией) данной конкретной формы, а не со статистическим ансамблем форм или классом изображений данной формы.

Мозаичные формы и их связь с CNN
(сегментация, генерация, обработка и анализ изображений, в том числе многоспектральных)

Результаты исследований по теме в рамках проекта НЦФМ

2022	• Введены *морфологии изображений (диффузная и проективная)* на основе мозаичных покрытий со всеми ранее известными инструментами сравнения форм и по форме.
2023	• Введен *формализм точечных покрытий*, который позволяет распространить теорию простоты и инструменты морфологического анализа на *широкий класс морфологий*, в том числе непрерывных и нечетких. • Введены *нечеткие модели сходства и классы форм* для описания несегментированных данных (изображений), позволяющие непосредственно оценивать их собственную простоту и совместную простоту с формами, за счет чего морфологическая корреляция изображений и форм может быть определена и вычислена без обращения к понятиям «оператора» и «проекции», исключительно инструментами теории простоты.
2024	• На основе меры простоты как *структурной вероятности* предложены соответствующие инструменты анализа сегментированных и несегментированных данных на основе *структурной информативности* (аналоги инструментов теории информации). • На основе модели точечных покрытий подход распространен на другие типы математических морфологий: морфологии Серра, Хаф-морфологии, структурные модели с направленными отношениями, тензорные модели отношений, смешанные мозаично-фигурные модели и т.д.

Визильтер Ю. В., Брянский С. А. Морфологическое сравнение форм и изображений на основе теории простоты и модели точечных покрытий, ВКИТ, 2024

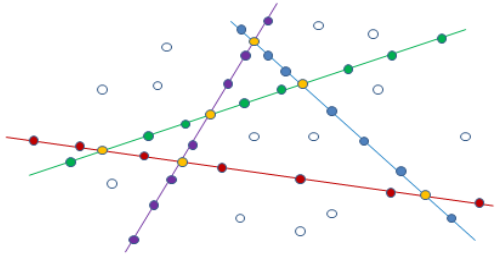
Результаты исследований по теме в рамках проекта НЦФМ

Предложены:

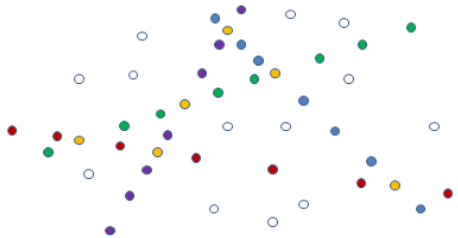
- **Формализм точечных покрытий**, который позволяет распространить теорию простоты и инструменты морфологического анализа на широкий класс морфологий, а именно:
  - в модель точечных покрытий вкладывается не только морфология Пытьева, но и ММ Серра и др., т.е. инструменты будут едиными для морфологий разных типов;
  - такой подход позволяет перейти от простоты формы в целом к простоте как мере на точках кадра – внутреннему распределению локальной точечной простоты;
  - такой подход дает возможность построения на основе этих распределений различных содержательных морфологических статистик;
  - такой подход это дает возможность ввести оценки простоты и сравнения по форме не только сегментированных, но и несегментированных данных;
- **Квад-нормальные (естественные) нечеткие модели сходства для описания несегментированных данных** (изображений), позволяющие оценивать их простоту и совместную простоту с формами;
- **Проективная морфология нестрогих полутоновых форм**, позволяющая сравнивать несегментированные изображения по форме;
- **Способы сравнения форм несегментированных изображений** на основе морфологических коэффициентов корреляции, а также апостериорных и условных оценок простоты без обращения к понятиям «оператора» и «проекции», исключительно инструментами теории простоты.

$$K_{MC}^2(g,F) = \| P_F g(x,y) - P_O g(x,y) \|^2 / \| g(x,y) - P_O g(x,y) \|^2 =$$
$$= \sum_{i=1,..,n} (q(g \cap F_i) - q(g)\, q(F_i)) / (p_i\, \mu(g)) = (K_M^2(g,F) - q(g)) / \mu(g)$$

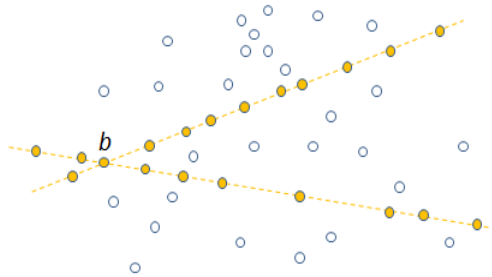
- Без проекторов
- Связь ММКц с МКК



Значимые прямые L_B



Разбиение F_{BL} (каждая желтая точка – отдельная область)



Точечное покрытие $\Phi_{BL}(b)$

Пример L-формы БТП
Покрытие любой точки – объединение всех точек, лежащих на проходящих через нее значимых прямых

Результаты исследований по теме в рамках проекта НЦФМ

Структурная энтропия

Простота точечного покрытия $\Phi_{(a)}$:

$q(\Phi) = \mathrm{mean}_{a\in\Omega}\, q(\Phi_{(a)})$.

Структурная энтропия $Y(\Phi)$:

$Y(\Phi) = \mathrm{mean}_{a\in\Omega^\Phi}\, Y(\Phi_{(a)}) = -\,\mathrm{mean}_{a\in\Omega^\Phi} \log_2 q(\Phi_{(a)})$,

$Y(\Phi_{(a)}) = -\log_2 q(\Phi_{(a)})$,

где $\Omega_\Phi = \{a\in\Omega\colon \Phi_{(a)} \neq \varnothing \Leftrightarrow q(\Phi_{(a)}) > 0\}$ – носитель формы Φ.

Взаимная структурная энтропия Γ и Φ :

$$Y(\Gamma \cap \Phi) = \mathrm{mean}_{a\in\Omega}\, Y(\Gamma_{(a)} \cap \Phi_{(a)}) = \\ = \mathrm{mean}_{a\in\Omega} \log_2 q(\Gamma_{(a)} \cap \Phi_{(a)}).$$

Условная структурная энтропия Γ относительно Φ:

$$Y(\Gamma \mid \Phi) = \mathrm{mean}_{a\in\Omega}\, Y(\Gamma_{(a)} \mid \Phi_{(a)}) = \\ = -\,\mathrm{mean}_{a\in\Omega} \log_2 q(\Gamma_{(a)} \mid \Phi_{(a)}).$$

***Основные свойства струкурной энотропии*:**

1. *Неотрицательность.*

$q(\Phi_{(a)})\in(0,1] \Rightarrow -\log_2 q(\Phi_{(a)}) \geq 0 \Rightarrow Y(\Phi) \geq 0$.

2. *Простейшая форма O имеет нулевую энтропию.*

$q(O_{(a)})\equiv 1 \Rightarrow -\log_2 q(O_{(a)}) = 0 \Rightarrow Y(O) = 0$.

3. *Точечное разбиение I имеет максимальную энтропию* $\log_2 S$.

$q(I_{(a)})\equiv 1/S \Rightarrow -\log_2 q(I_{(a)}) = \log_2 S \Rightarrow Y(I) = \log_2 S$.

4. *Индикатор идентичности.*

Если $\Phi=\Gamma$, то $Y(\Phi) = Y(\Gamma)$.

5. *Аддитивность для независимых форм:*

$Y(\Gamma \cap \Phi) = Y(\Gamma) + Y(\Phi)$.

6. *Аддитивная связь взаимной и условной энтропии:*

$Y(\Gamma \cap \Phi) = Y(\Gamma \mid \Phi) + Y(\Phi) = Y(\Phi \mid \Gamma) + Y(\Gamma)$.

Таким образом, $Y(\Phi \mid \Gamma)$ выражает дополнительную *структурную информацию*, которую форма Φ привносит в $\Gamma\cap\Phi$ дополнительно к исходной *структурной информации*, содержавшейся в Γ.

Результаты исследований по теме в рамках проекта НЦФМ

МККФ	УСЭ
а) сравнение с простейшей формой дает простоту $K_M^2(\Phi,O) = q(\Phi)$;	*а) сравнение с простейшей формой дает энтропию* $Y(\Phi \mid O) = Y(\Phi)$;
б) асимметричность $K_M^2(\Gamma,\Phi) \neq K_M^2(\Phi,\Gamma)$;	*б) асимметричность* $Y(\Gamma \mid \Phi) \neq Y(\Phi \mid \Gamma)$;
в) индикатор тождества форм $K_M^2(\Phi,\Phi) = 1$;	*в) индикатор тождества форм* $Y(\Phi \mid \Phi) = 0$;
г) индикатор упорядоченности форм по сложности $K_M^2(\Phi,\Gamma) = 1 \Leftrightarrow \Gamma \subseteq \Phi$; $K_M^2(\Phi,\Gamma) = K_M^2(\Gamma,\Phi) = 1 \Leftrightarrow \Gamma = \Phi$;	*г) индикатор упорядоченности форм по сложности* $Y(\Phi \mid \Gamma) = 0 \Leftrightarrow \Gamma \subseteq \Phi$; $Y(\Phi \mid \Gamma) = Y(\Gamma \mid \Phi) = 1 \Leftrightarrow \Gamma = \Phi$;
д) индикатор независимости форм $K_M^2(\Gamma,\Phi) = 0 \Leftrightarrow K_M^2(\Phi,\Gamma) = 0 \Leftrightarrow$ $\Leftrightarrow q(\Gamma \cap \Phi) = q(\Gamma)\, q(\Phi)$.	*д) индикатор независимости форм* $Y(\Gamma \mid \Phi) = Y(\Gamma) \Leftrightarrow Y(\Phi \mid \Gamma) = Y(\Phi) \Leftrightarrow$ $\Leftrightarrow q(\Gamma \cap \Phi) = q(\Gamma)\, q(\Phi)$.

Сравнение условной структурной энтропии (УСЭ) как *среднего минус логарифма* условной простоты с морфологическим КК форм (МККФ) как *средней* условной простотой:

$$Y(\Gamma \mid \Phi) = -\,\mathrm{mean}_{a\in\Omega} \log_2 q(\Gamma_{(a)} \mid \Phi_{(a)})$$

$$K_M^2(\Gamma,\Phi) = -\,\mathrm{mean}_{a\in\Omega}\, q(\Gamma_{(a)} \mid \Phi_{(a)}).$$

Таким образом, **условная структурная энтропия (УСЭ) может использоваться в качестве функционального аналога морфологического коэффициента корреляции форм (МККФ) в различных задачах сравнения форм во всех описанных выше типах морфологий**, в которых модели формы описываются в виде покрытий.

Результаты и задачи дальнейших исследований по теме в рамках проекта НЦФМ

Отношения более 2 элементов

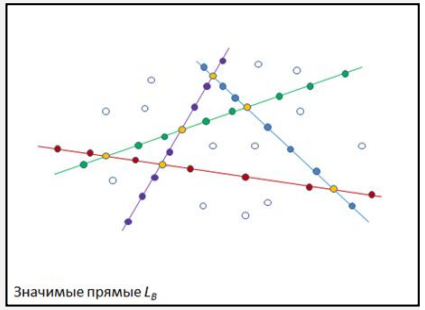
t-простота и фрактальная t-энтропия

t-простота для случая произвольных мультиплетов (*t-плетов*): геометрическая вероятность того, что t случайных точек БТП A (*t-плет*) попадут на одну прямую из L_A:

$$q_{(t)}(\Phi_{AL}) = \mathrm{mean}_{a\in A}\, q_{a(t)}(\Phi_{AL}) = \mathrm{mean}_{a\in A} \sum_{i=1,..,n(t)} 1(a\in l_{A(t)i})\, z_{\Phi AL(t)i}^{(t-1)}.$$

В частности, при отсутствии пересечения t-значимых прямых в A:

$$q_{(t)}(F_{AL}) = \mathrm{mean}_{a\in A}\, q_{a(t)}(F_{AL}) = \sum_{i=1,..,n(t)} p_{FAL(t)i}\, q_{a(t)}(F_{AL}) =$$
$$= \sum_{i=1,..,n(t)} p_{FAL(t)i}\, p_{FAL(t)i}^{(t-1)} = \sum_{i=1,..,n(t)} p_{FAL(t)i}^{t}.$$



Структурная t-энтропия $Y_{(t)}(\Phi_{AL})$:

$$Y_{(t)}(\Phi_{AL}) = \mathrm{mean}_{a\in A}\, Y_{a(t)}(\Phi_{AL}) = -\,\mathrm{mean}_{a\in A} \log_2 q_{a(t)}(\Phi_{AL}).$$

В частности, при отсутствии пересечения t-значимых прямых в A:

$$Y_{(t)}(F_{AL}) = \mathrm{mean}_{a\in A}\, Y_{a(t)}(F_{AL}) = \sum_{i=1,..,n(t)} p_{FAL(t)i}\, Y_{a(t)}(F_{AL}) =$$
$$= (t-1)\, Y_{(2)}(F_{AL}) = (t-1)\, Y(F_{AL}).$$

t-инвариантной структурной t-энтропией $Y^*_{(t)}(\Phi_{AL})$ назовем величину:

$$Y^*_{(t)}(\Phi_{AL}) = Y_{(t)}(\Phi_{AL}) / (t-1) = (-\,\mathrm{mean}_{a\in A} \log_2 q_{a(t)}(\Phi_{AL}))\,/\,(t-1). \quad (13.11)$$

Это напоминает *фрактальную размерность*, которая инвариантна к масштабу измерения. Такую оценку можно также назвать *фрактальной t-энтропией*.

Основная задача исследований: *научиться вычислять кросс-модальную морфологическую корреляцию и структурную взаимную/условную энтропию для данных разных модальностей (и разной структуры), если они покрывают один предметный домен*

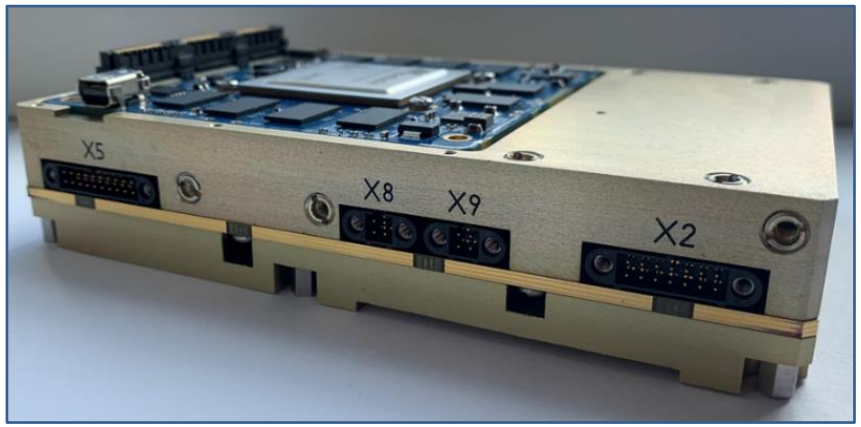
Например, мы планируем морфологически сравнивать изображение сцены и его словесное описание

Анализ и синтез многоспектральных изображений для САУ ЛА

Подр. В.А.Князя, Лаб. Д.В.Комарова, Лаб. В.В.Князя, Лаб. В.А.Горбачева

Бортовая вычислительная платформа с нейроускорителем

В целях использования в автономных аппаратах воздушного и подводного базирования кооперацией ГосНИИАС созданы экспериментальные образцы **унифицированных** бортовых аппаратно-программных вычислительных платформ. Платформа построена на базе **отечественной ЭКБ**.



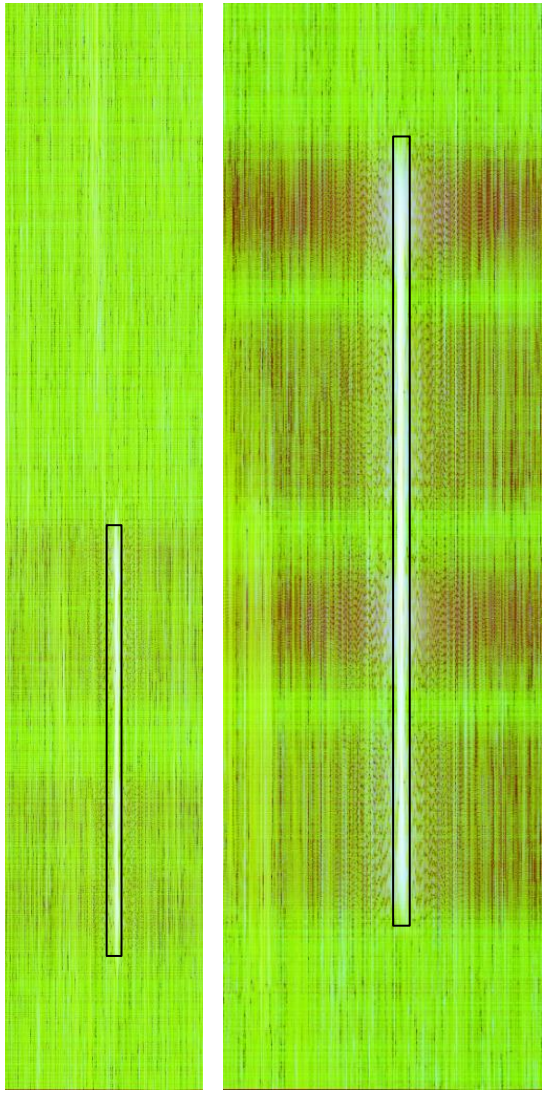
Центральным элементом платформы является **нейромезонин** – специальный вычислитель, обеспечивающий реализацию нейронных сетей, решение задач цифровой обработки сигналов и изображений **(на базе NM6408 от НТЦ «Модуль»)**

Создание виртуальных сцен

Пример сцены в видимом диапазоне

Пример сцены в инфракрасном диапазоне

Пример из набора гидроакустических данных



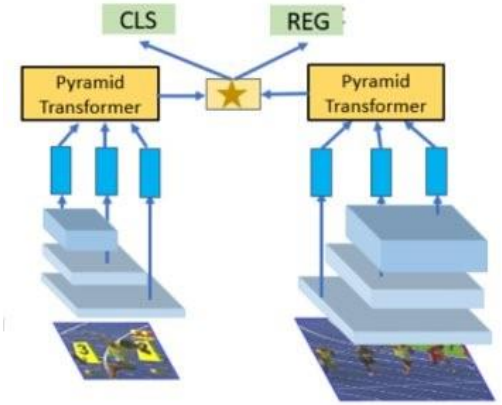
Методы нейросетевого поиска объекта на изображении по шаблону

Исследованы методы нейросетевого поиска объекта на изображении по заданному шаблону. Рассматривались **методы SiamMask и SiamTPN**. Метод SiamTPN показал принципиальную применимость для данной задачи.

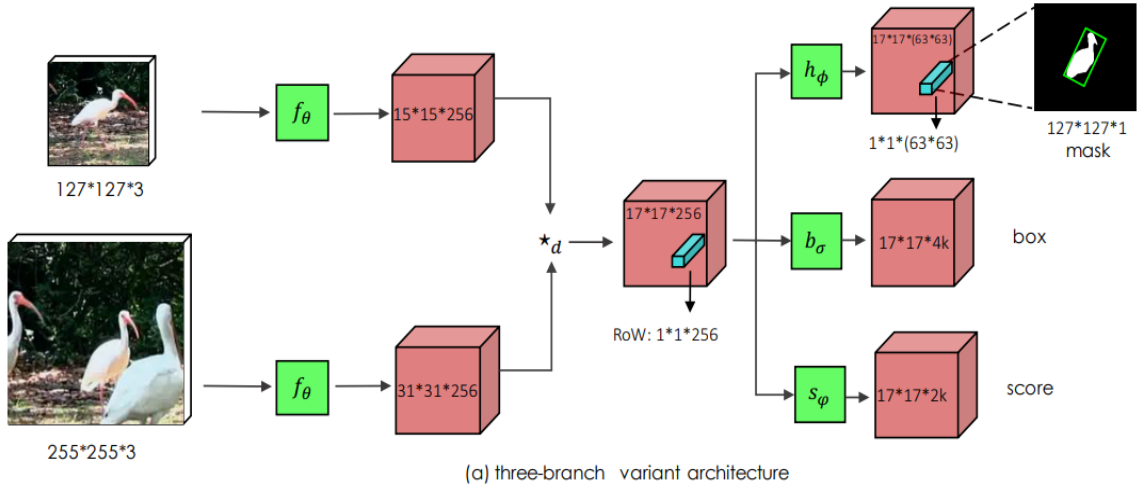
Синтетический набор данных

Результаты тестов

	Success	Precision
SiamTPN	0.730	0.939
SiamMask	0.674	0.781



Архитектура SiamTPN



Архитектура SiamMask

Методы автоматического машинного обучения для задачи многоспектрального детектирования объектов

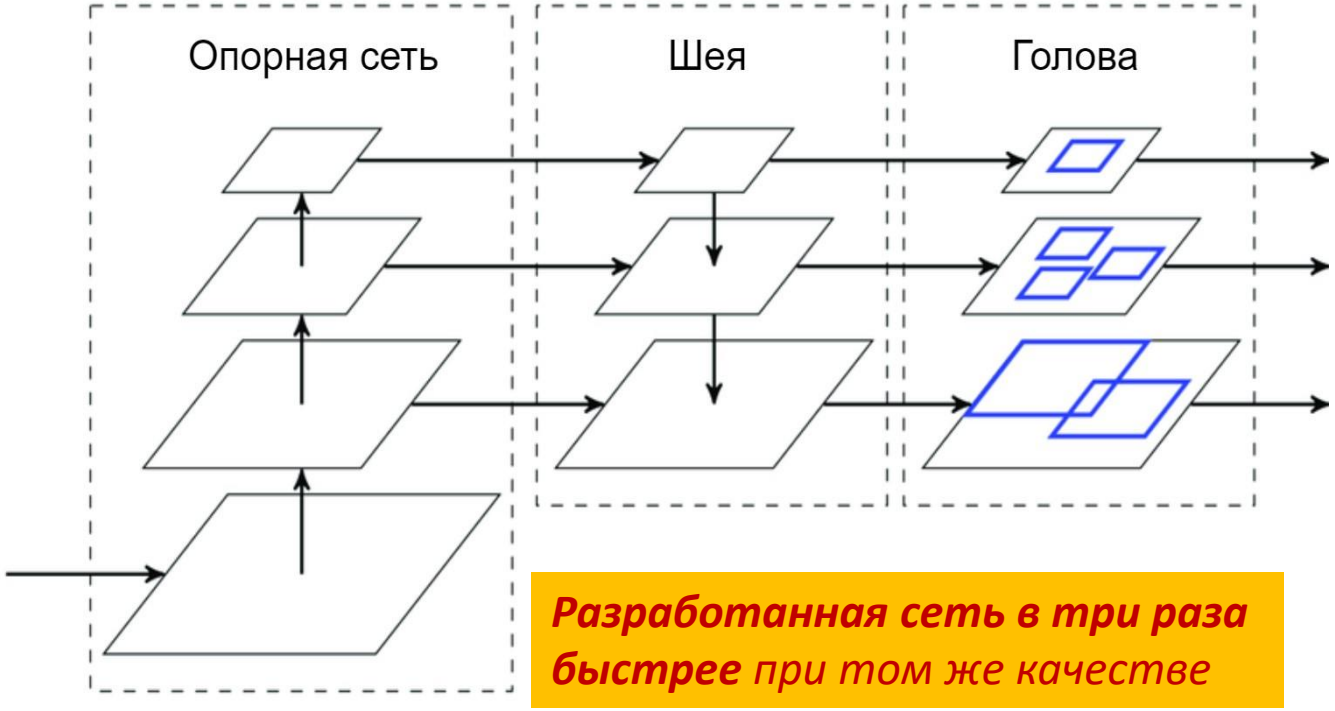
За основу взят алгоритм автоматического формирования архитектуры Once-for-All (OFA):

1. Обучение «суперсети» с вариативными параметрами (глубина, количество каналов и размер сверточного ядра)
2. Эволюционный поиск оптимальной «подсети»

- **Архитектура суперсети** – последовательность сверточных блоков с настраиваемыми параметрами (каждом блоке автоматически подбирается глубина, количество каналов и размер сверточного ядра)
- **Автоматическое формирование архитектуры** при ограничении на время работы в задаче классификации

В оригинальный алгоритм OFA добавлены новые модули для решения задачи детектирования объектов:

- Сформирована составная сеть, состоящая из опорной сети, шеи и головы детекции
- Опорная сеть – исходная подбираемая OFA-архитектура для извлечения карт признаков разного масштаба
- Шея – модуль на базе FPN с подбираемыми параметрами для эффективного детектирования объектов на разных масштабах
- Голова – YOLO-модуль для предсказания и декодирования классов и координат объектов



Тестирование на базе реальных данных

Разработанная сеть в три раза быстрее *при том же качестве*

Тестирование на процессоре NM6408

Проект «САПФИР»

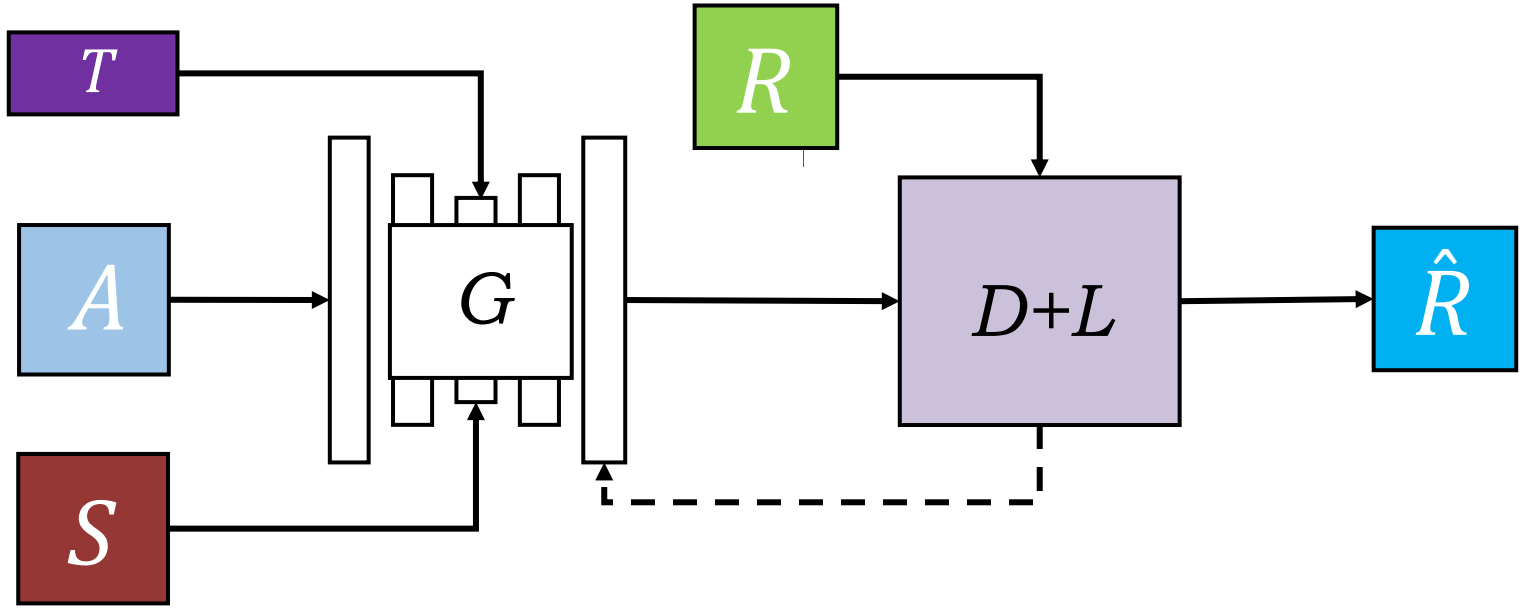
- **Разработка технологии реалистичного моделирования изображений дальнего инфракрасного диапазона (8-14 мкм)**
- Отличительные особенности СПК:
  - Статистически обоснованное повторение характеристик реальных датчиков на основе машинного обучения
  - Параметрическое управление погодными условиями, задаваемое одним эталонным изображением инфракрасного диапазона

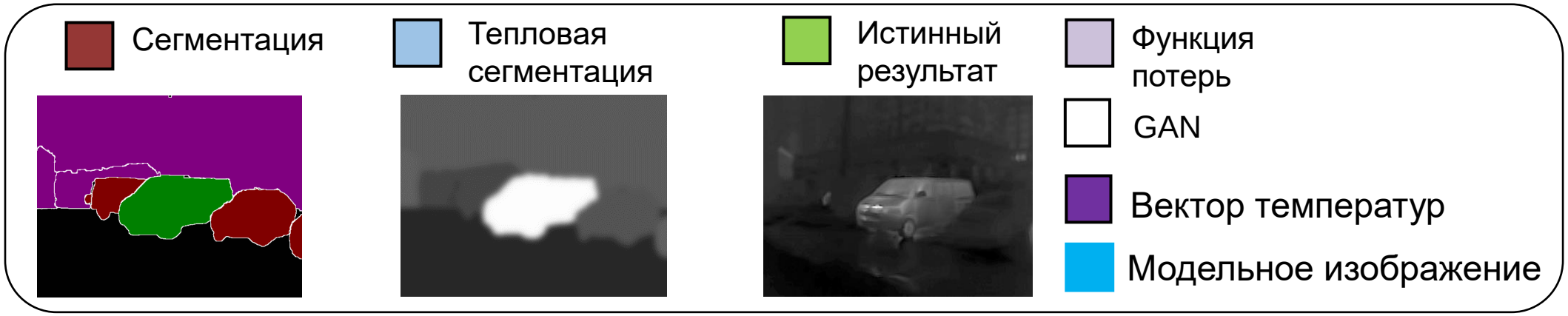
Реальная съёмка из лётного эксперимента

1. **Анализ и разметка реальных данных**
2. **Обучение нейросетевых моделей излучения объектов, атмосферы и датчика**
3. **Синтез модельных изображений**

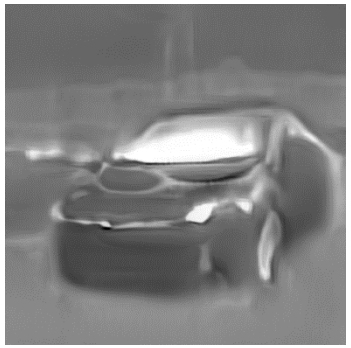
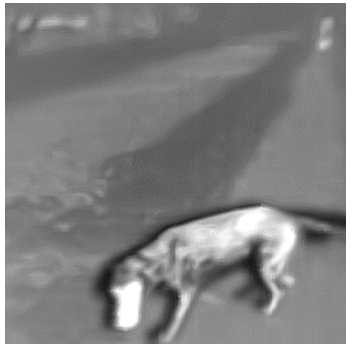
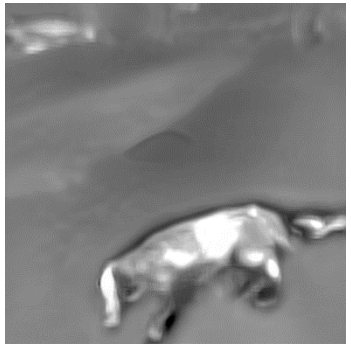
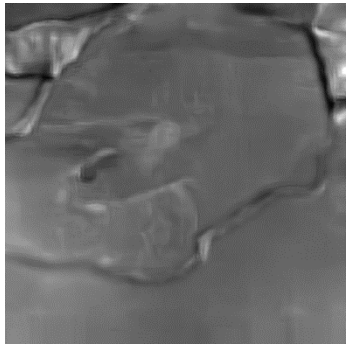
СПК Сапфир — СПО Сцена, результат моделирования

Постановка задачи синтеза ИК изображений из изображений видимого диапазона

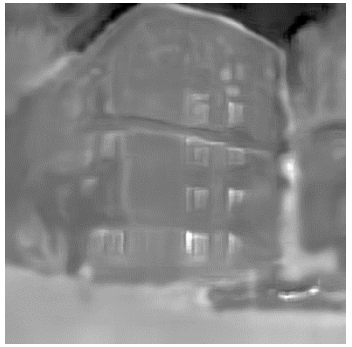
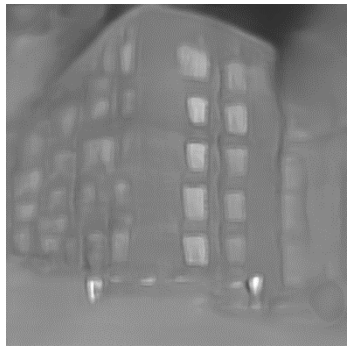
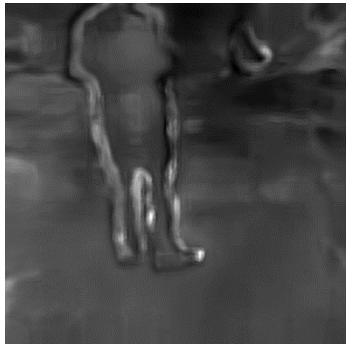
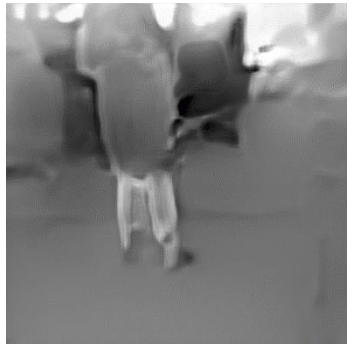
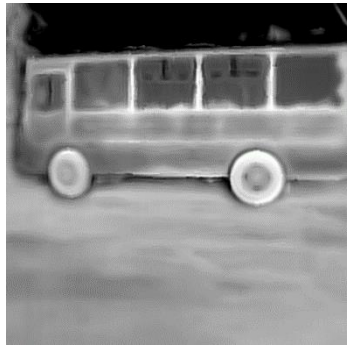
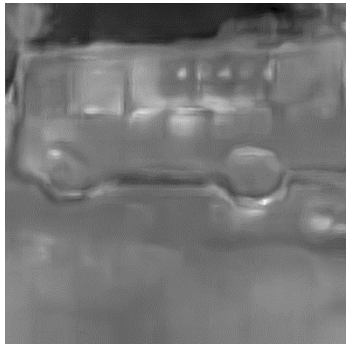
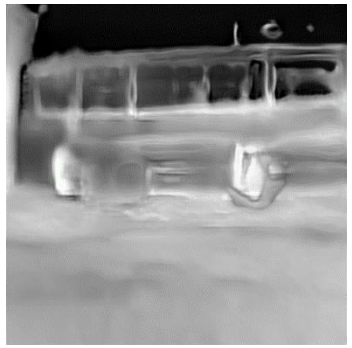




Примеры синтеза изображений из ТВ в ИК

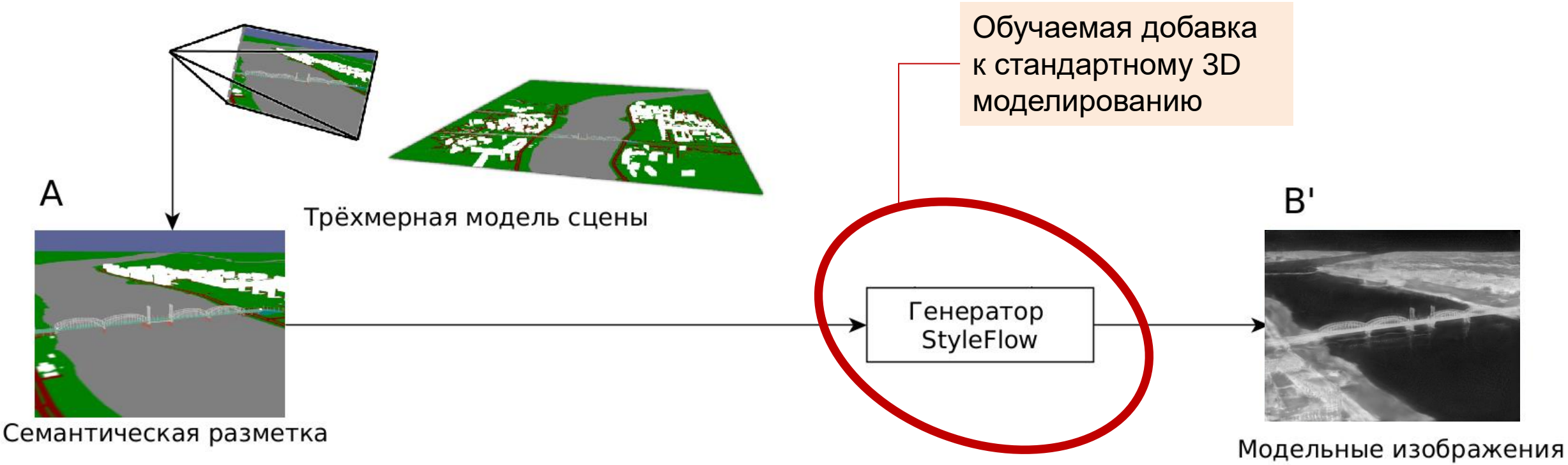
Вход	Истиное изображение	Pix2Pix	ThermalGAN	Reduce LayerGAN	LayerGAN
					
					
					

Примеры синтеза изображений из ТВ в ИК

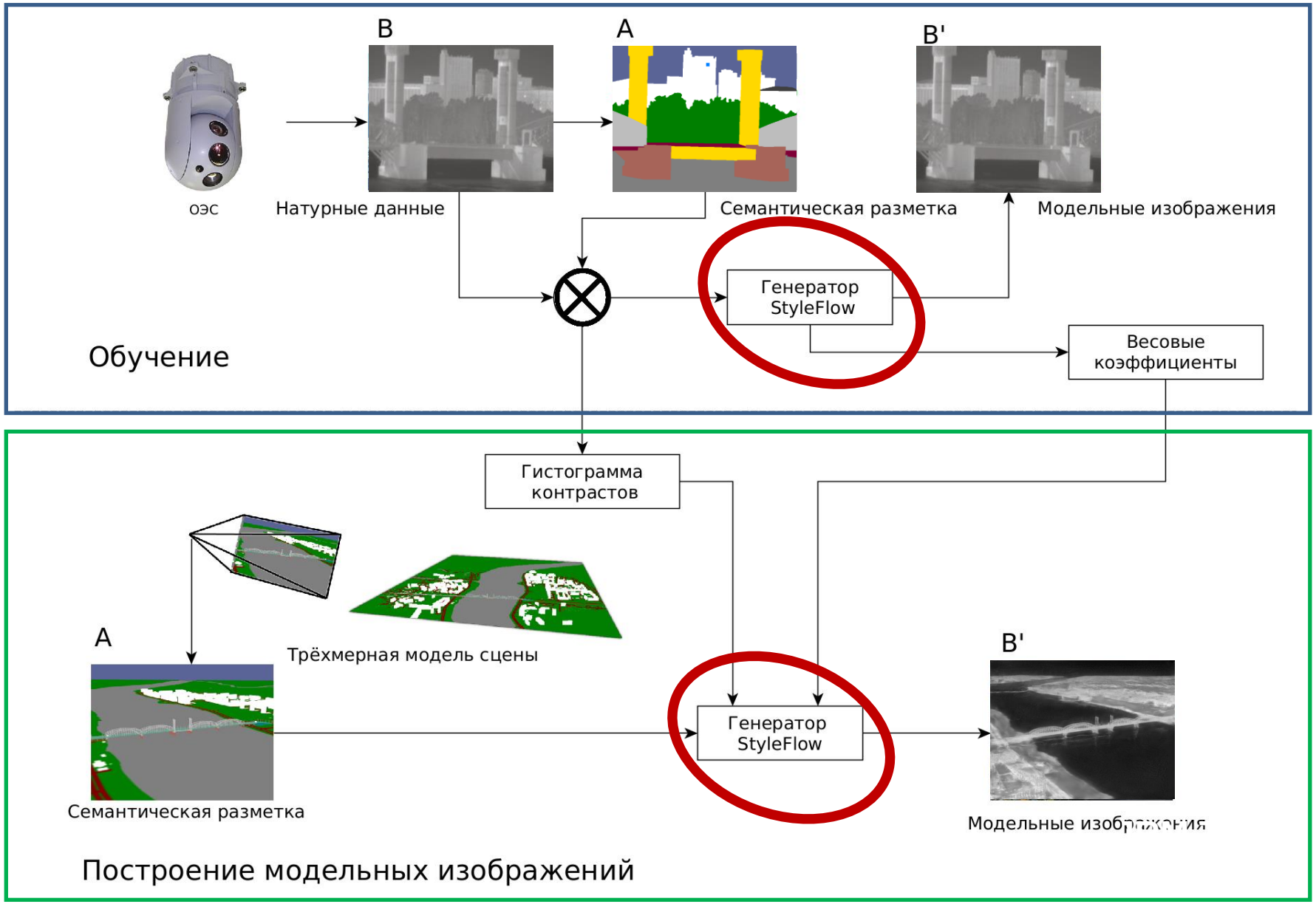
Вход	Истиное изображение	Pix2Pix	ThermalGAN	Reduce LayerGAN	LayerGAN
					
					
					

Постановка задачи синтеза ИК изображений из семантической 3D модели сцены

- Рассматривается задача в следующей постановке: дана трёхмерная модель сцены ***S*** с текстурами, соответствующими семантическим классам объектов (трава, дерево, здание и т. п.), даны геометрические параметры камеры, дан небольшой объём натурных данных для произвольной сцены, содержащей аналогичные классы объектов ***B*** и их семантическая разметка ***A***.
- Требуется разработать метод синтеза модельных изображений *f(**A**)*, формирующий модельные изображения ***B'*** максимально точно соответствующие реальным изображениям ***B***.



Функциональная схема метода нейросетевого моделирования



Синтез многоспектральных изображений для отработки и обучения САУ ЛА

Сравнение с аналогами

Распределённые многоспектральные сенсоры на основе БЛА и РТК

Лаб. В.А.Горбачева, Подр. Б.В.Вишнякова

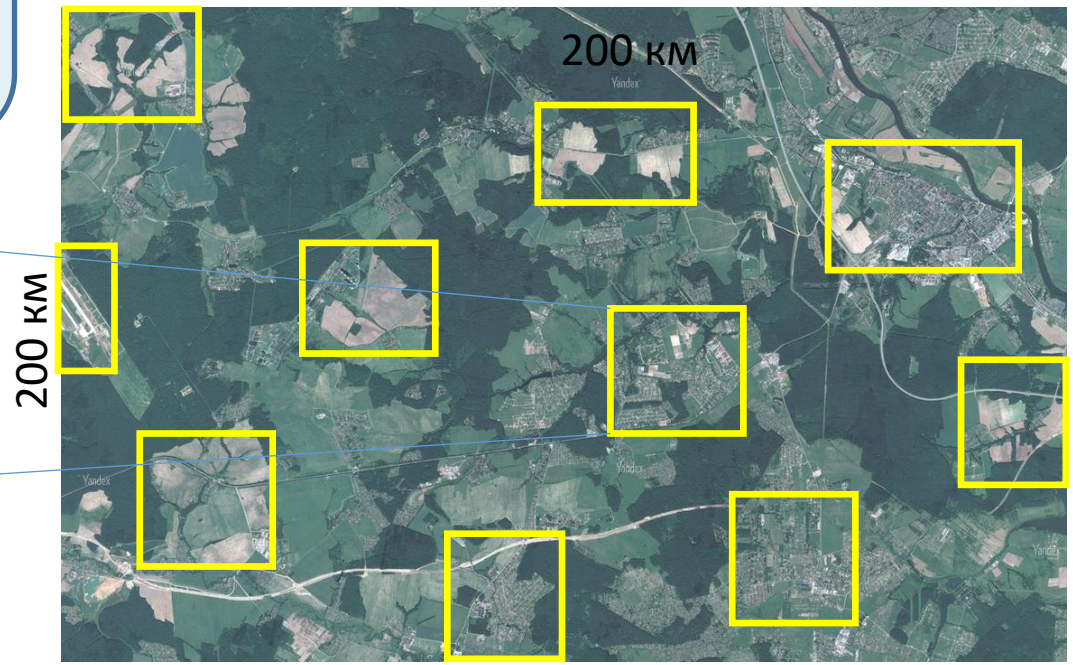
Комплекс распределённого автоматизированного многоспектрального мониторинга на базе БЛА

- **ТВ**
  - Высокое разрешение
  - Высокое качество в светлое время суток
- **ИК**
  - Повышение информативности в ограниченно сложных условиях наблюдения
- **РЛ**
  - Всепогодность, всесуточность
  - Более низкое разрешение (1 м/пикс)

Принцип многоспектральности: *наблюдение объектов в 3 различных диапазонах для максимизации вероятности обнаружения*

- **Звено** из 4 БЛА с разноспектральными ЦН
- Фиксированные задачи БЛА
- Одна стартовая позиция

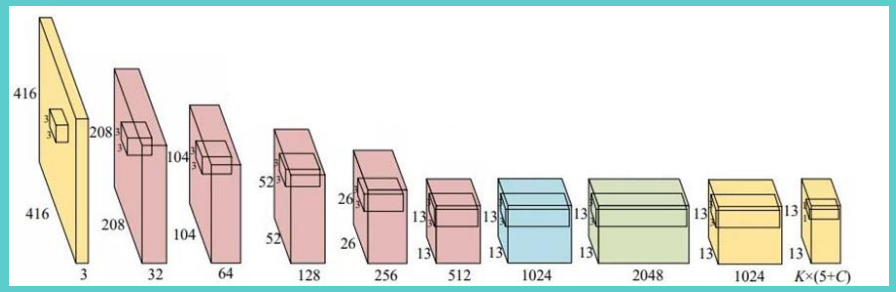
- **Несколько звеньев** БЛА
- Перераспределение задач и аппаратов
- Несколько стартовых позиций



Комплекс распределённого автоматизированного многоспектрального мониторинга на базе БЛА

Разработка облика и изготовление комплекса на базе группы БЛА

Разработка и изготовление бортового спецвычислителя

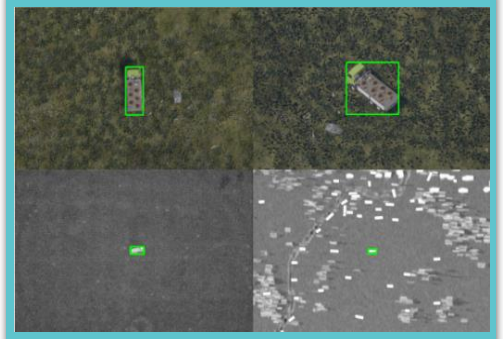


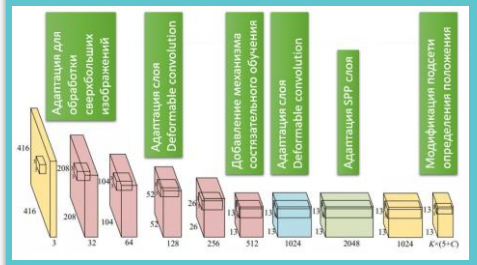
Разработка нейросетевых алгоритмов обработки данных и бортового ПО

Создание и наполнение БД реальных обучающих изображений

Алгоритмы обработки информации и управления на борту БЛА

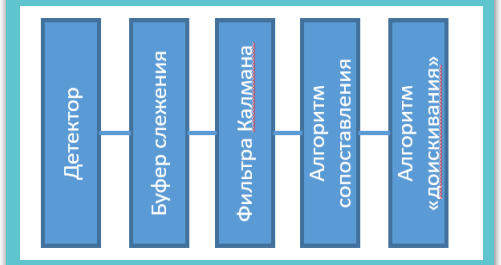
Алгоритмы обнаружения и распознавания





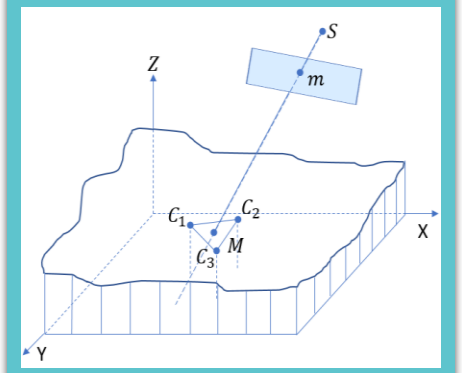
Обнаружение и распознавание объектов на данных различных спектральных диапазонов

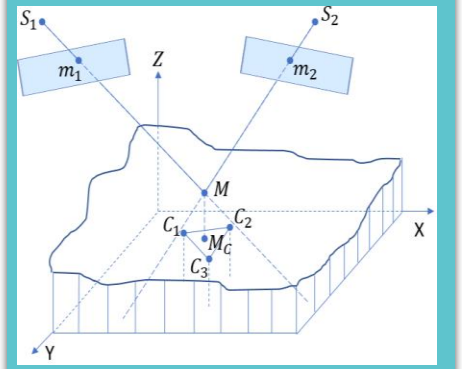
Алгоритмы трекинга



Межкадровое сопоставление объектов

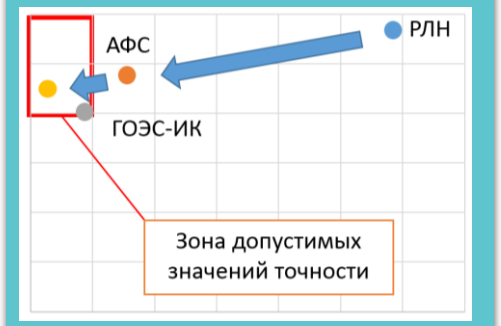
Алгоритмы геопривязки

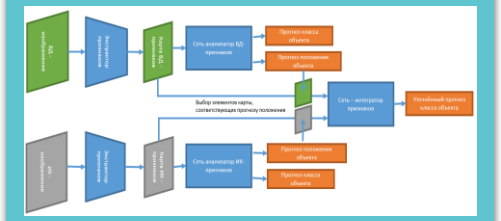




Алгоритмы определения положения объекта по одному, двум и более снимкам объекта

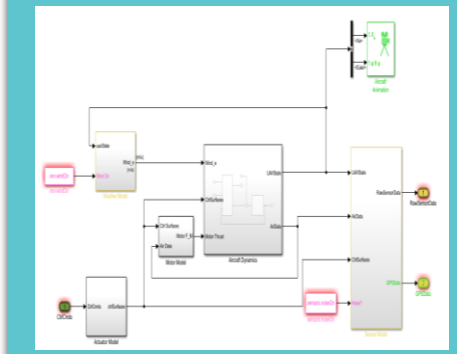
Алгоритмы комплексирования

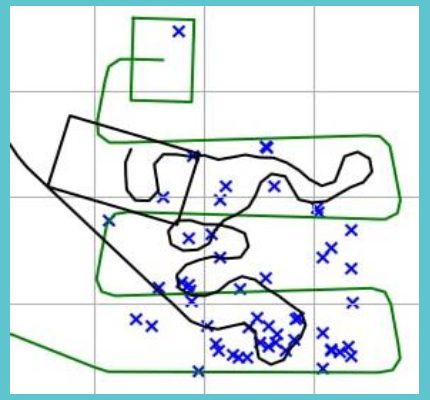




Алгоритм сопоставления объектов и комплексирования данных от распределённых датчиков

Алгоритмы управления



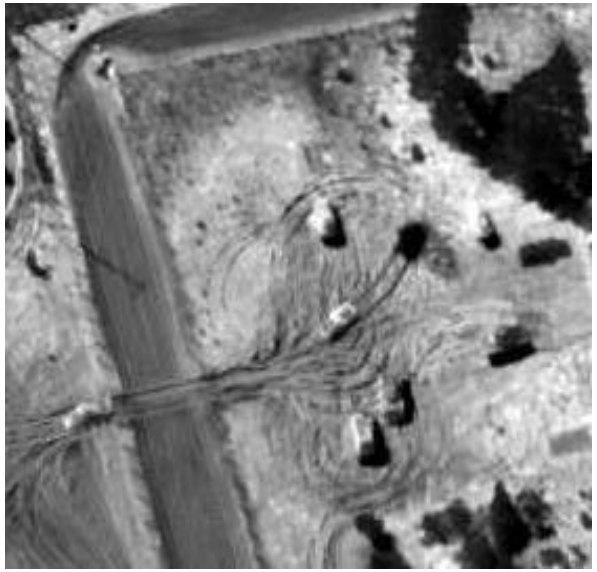


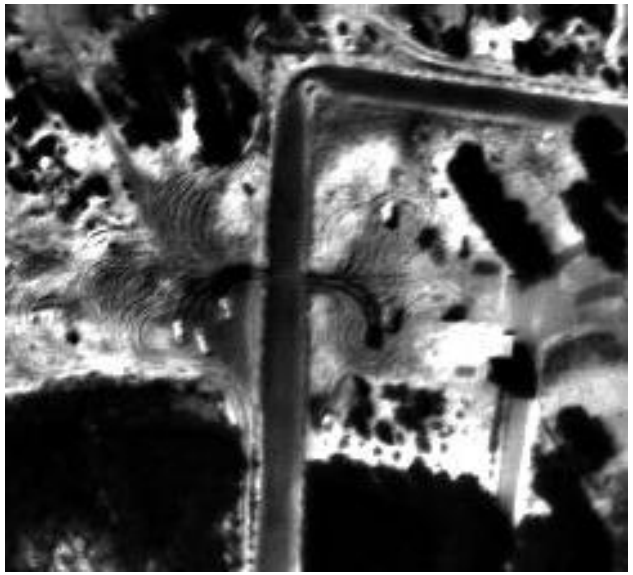
Алгоритмы для построения оптимальных маршрутов БЛА

Пример изображений полученных с датчиков БЛА

Обзорная и детальная телевизионная съёмка

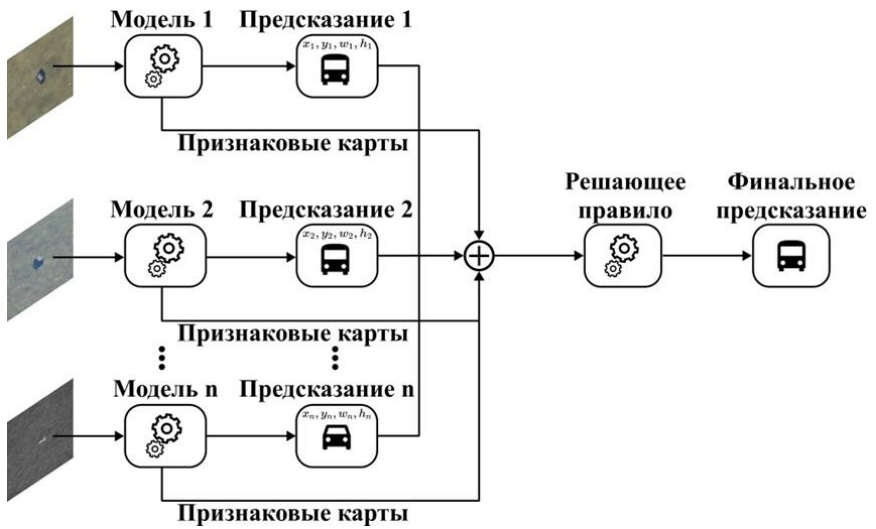
Тепловизионная съёмка





РЛ-наблюдение

Нейросетевой метод комплексирования разновременных разноспектральных изображений от распределенного сенсора



Ансамблирование моделей на уровне признаковых пространств

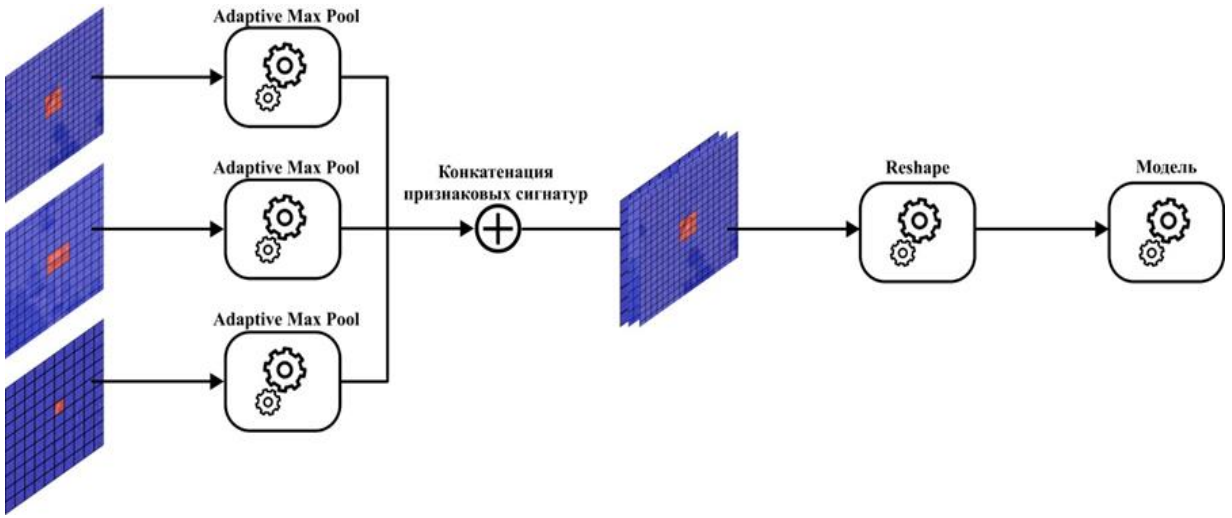
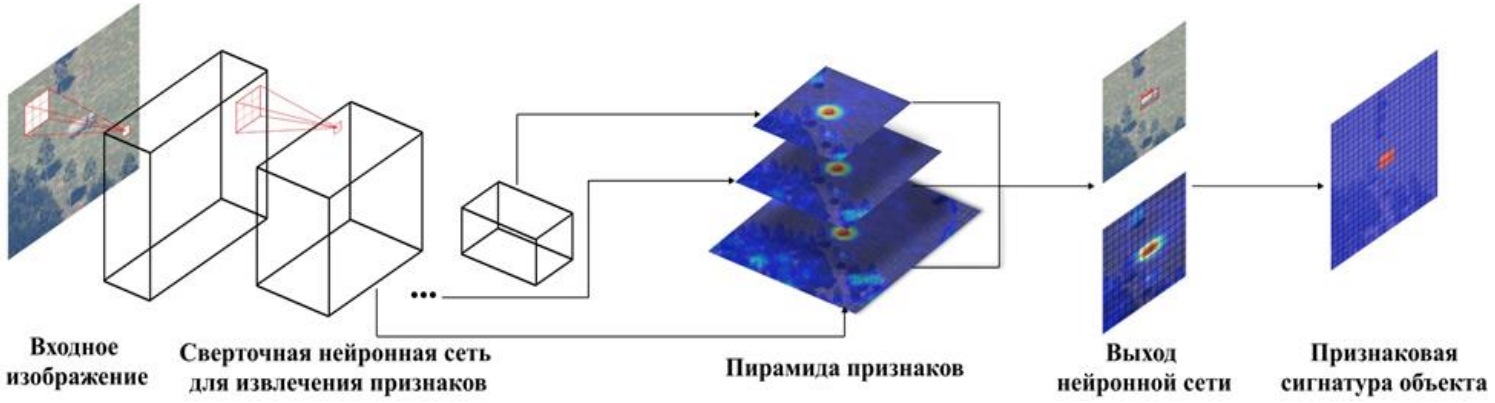


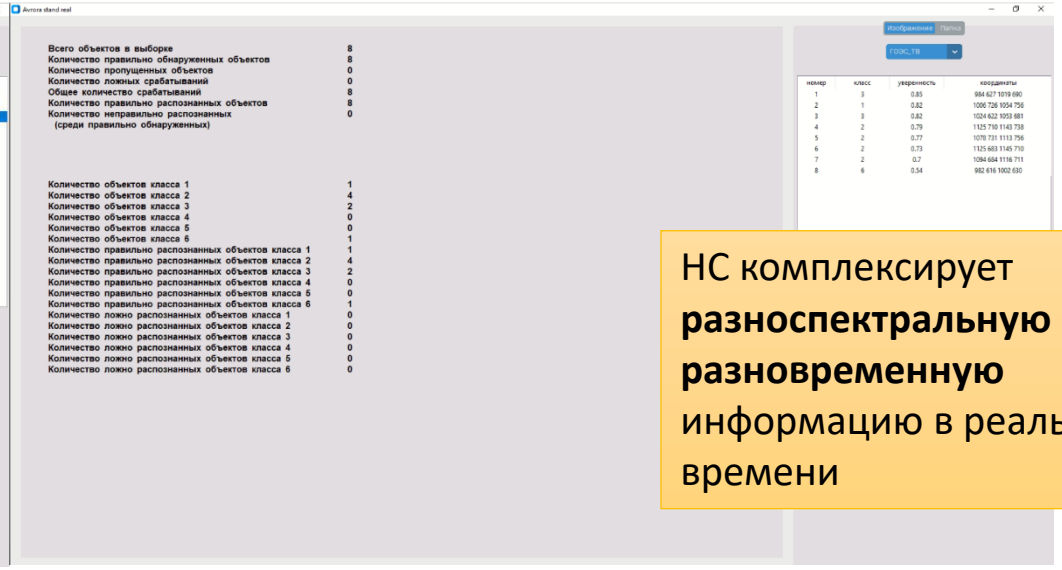
Схема слияния признаковых представлений объекта



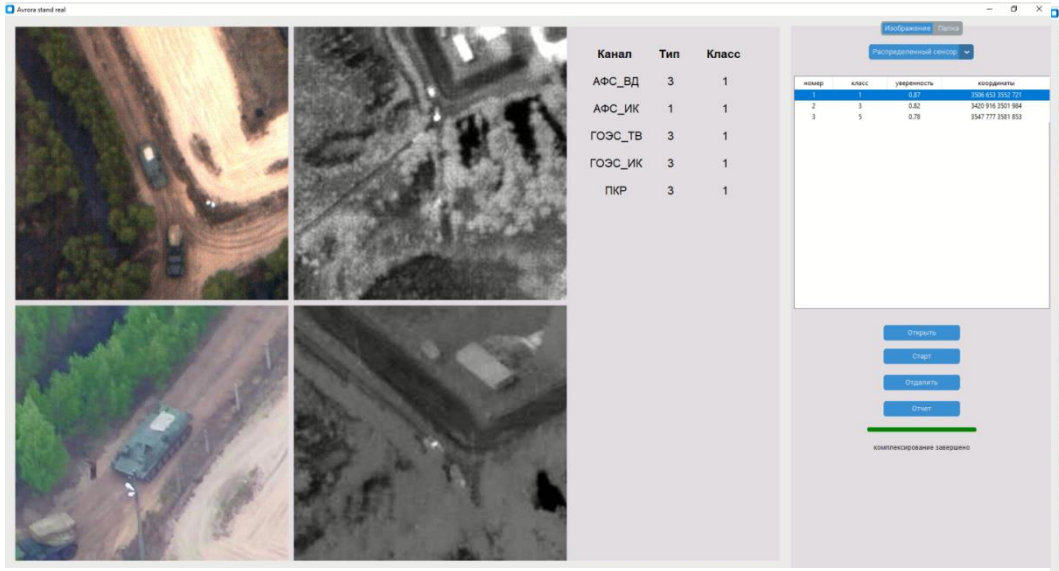
Формирование признаковой сигнатуры объекта

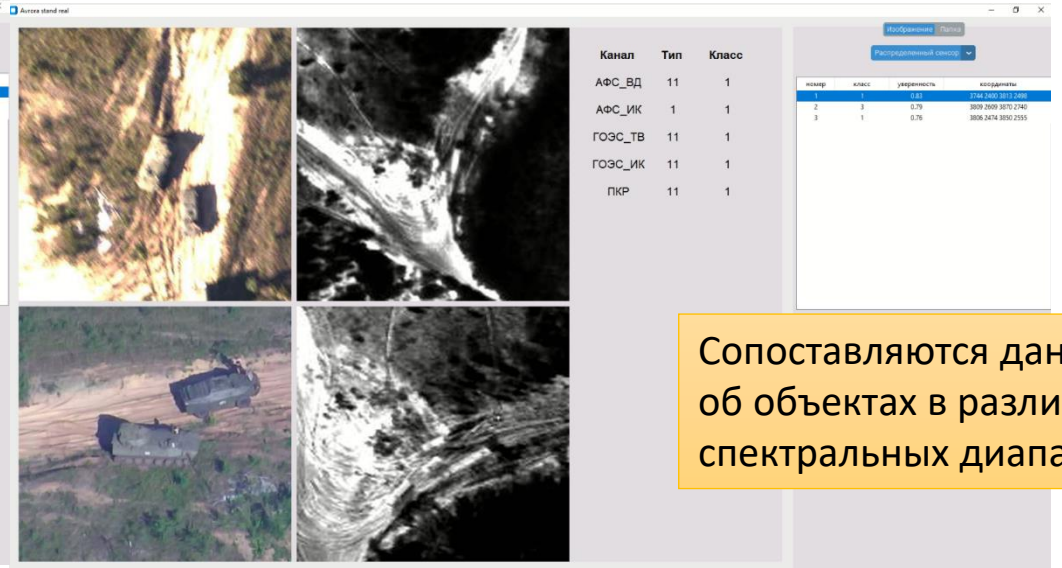
Методы на основе ансамблирования моделей могут обрабатывать данные, получаемые с различных ракурсов. Однако, при объединении предсказаний нейросетевых моделей теряется вся признаковая информация об объекте наблюдения. Поэтому **предложен подход, при котором комплексирование происходит на уровне признаковых представлений**. Для этого первоначально извлекаются признаковые карты и подаются на вход алгоритму, который их обрабатывает и делает финальное предсказание

Принятие комплексного решения об объекте



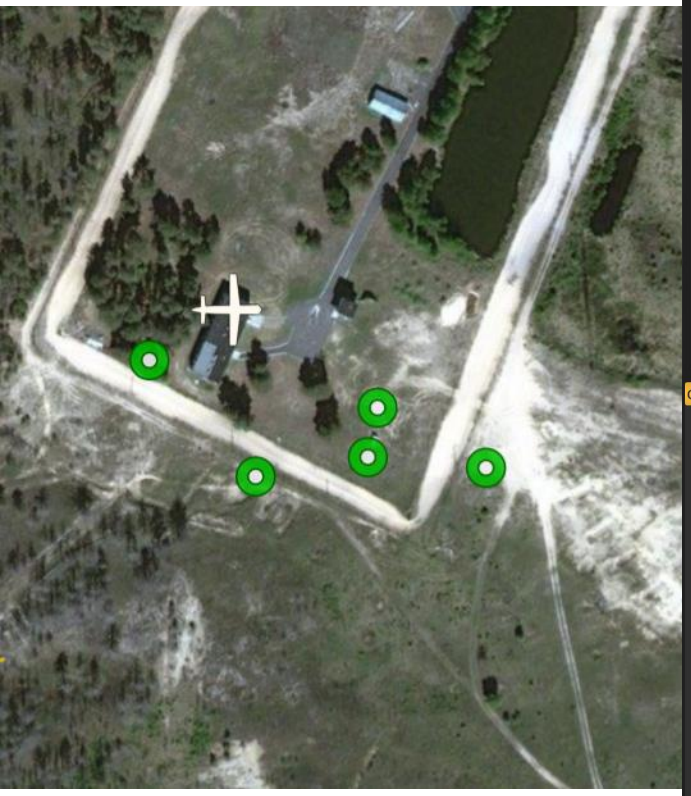
НС комплексирует **разноспектральную разновременную** информацию в реальном времени





Сопоставляются данные об объектах в различных спектральных диапазонах

Принятие комплексного решения об объекте



Бортовое ПО БЛА автоматически готовит и передаёт на наземный пункт управления оператору **формуляры обнаруженных объектов**:

- Изображение
- Класс
- Тип
- Координаты

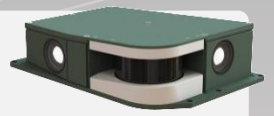
ОТКРЫТАЯ МОДУЛЬНАЯ АРХИТЕКТУРА ПОСТРОЕНИЯ СТЗ РТК

Оптико-электронный приборный блок

- Камера ТВ диапазона
- Камера ИК диапазона
- Дальномер

Блок формирования трехмерной модели местности

- Лидар
- Широкоугольные ТВ камеры

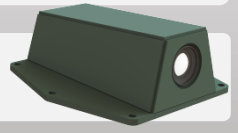


Фронтальный блок камер дальнего обзора

- Камера коротковолнового ИК диапазона
- Узкоугольная ТВ камера

Камера бокового обзора

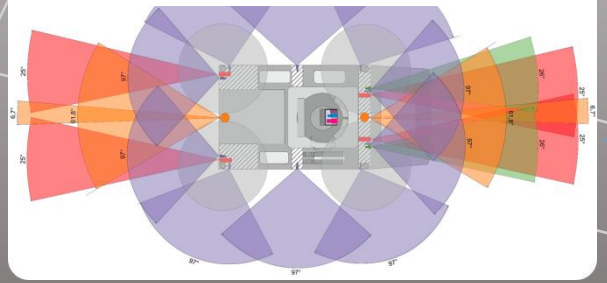
- Широкоугольная ТВ камера



Фронтальный радар

- Многорежимный электронный сканирующий радар

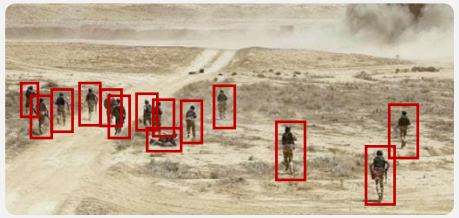
Поле зрения датчиков сенсорной системы РТП

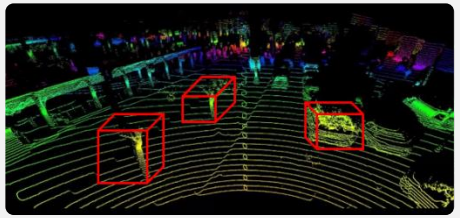


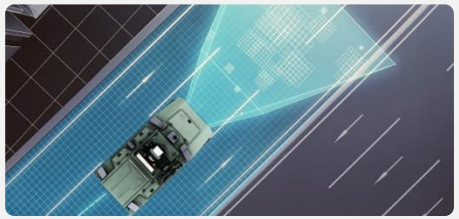
Бортовые вычислители многопоточной обработки данных

- Обнаружение и распознавание целей
- Прослеживание (ведение) целей
- Определение точки прицеливания

- Поиск и распознавание препятствий
- Построение карты проходимости







- Построение трехмерной модели местности
- Интеграция в общую информационную модель

- Семантическая сегментация сцены

ПРОЕКТ «МАРКЕР»

Создание экспериментальной автономной робототехнической платформы для отработки ключевых технологий наземной робототехники

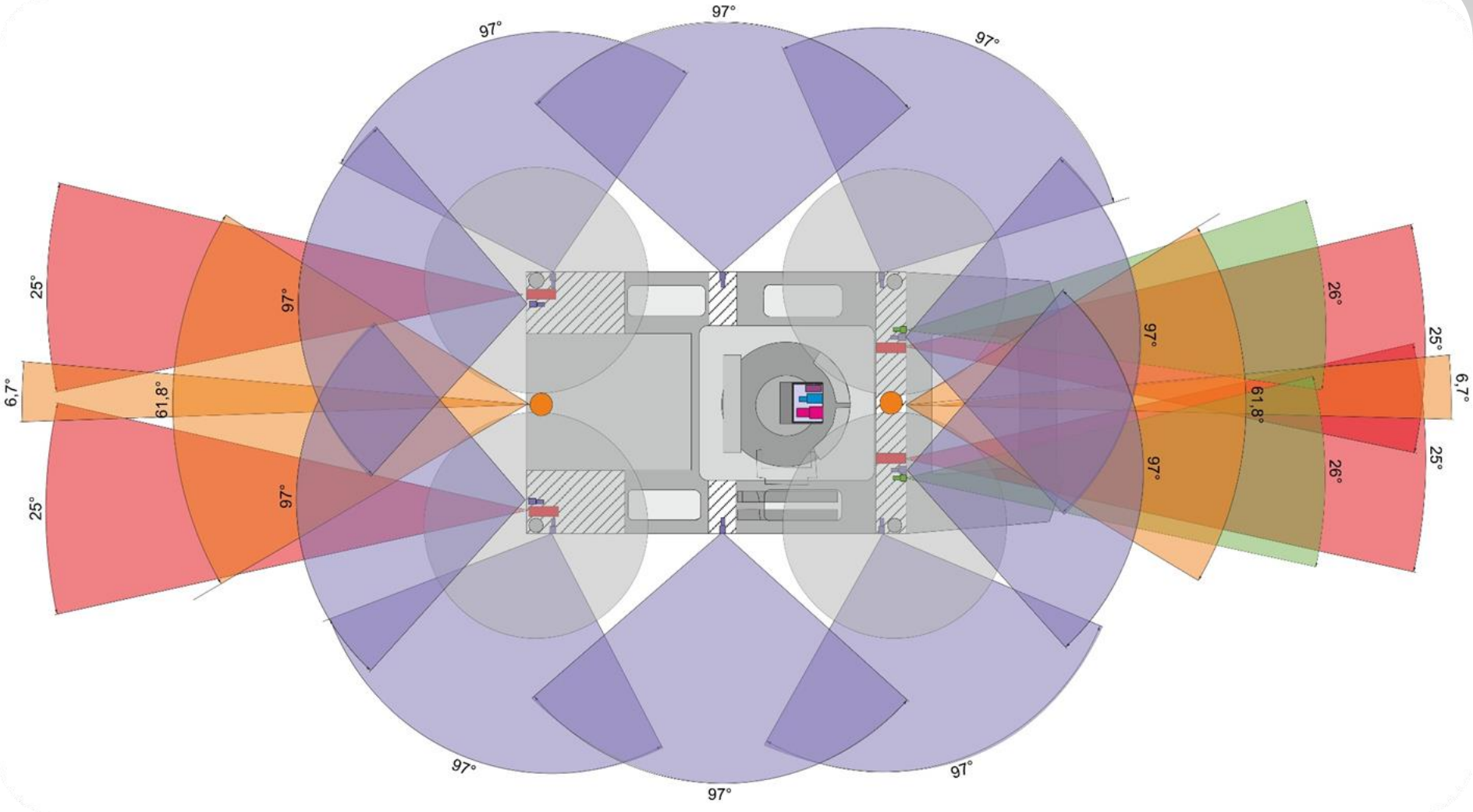
СТЗ РТП, ПОСТРОЕННАЯ ПО МОДУЛЬНОМУ ПРИНЦИПУ

Датчики РТП

- Тепловизоры
- Лидары
- Камеры PTZ
- Комплект камер полноразмерного обзора
- Комплекты теле стереокамер
- Радары

Датчики ОЭПБ

- Прицельная камера
- Тепловизор
- Дальномер



ТЕХНОЛОГИЯ ОПЕРАТИВНОГО ДООБУЧЕНИЯ АЛГОРИТМОВ СТЗ

ТЕХНОЛОГИЯ ОПЕРАТИВНОГО ДООБУЧЕНИЯ АЛГОРИТМОВ СТЗ

МЕТОДЫ ПОСТРОЕНИЯ СЕМАНТИЧЕСКОЙ ТРЕХМЕРНОЙ МОДЕЛИ СЦЕНЫ

Разработанный интегральный алгоритм

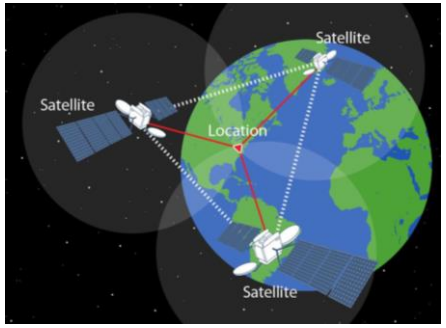
Входные данные

Данные с датчиков:

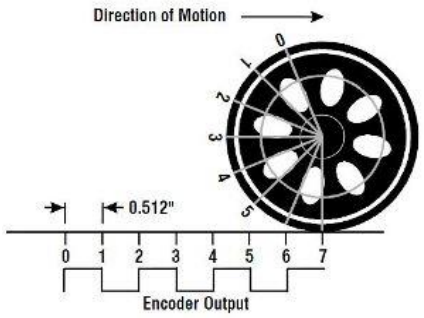
Стереокамеры

Лидары

GPS



Датчики углов (IMU)



Алгоритм трехмерной реконструкции

Составление трехмерного плотного облака точек

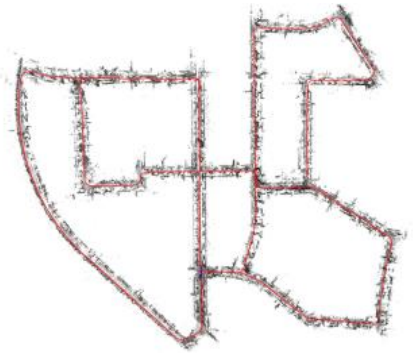
Определение местоположения РТП в пространстве (одометрия)

Алгоритм семантической сегментации

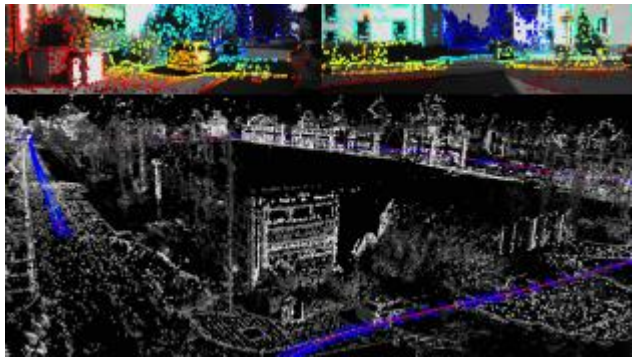
«Раскрашивание» облака точек в цвета сегментов

Выходные данные

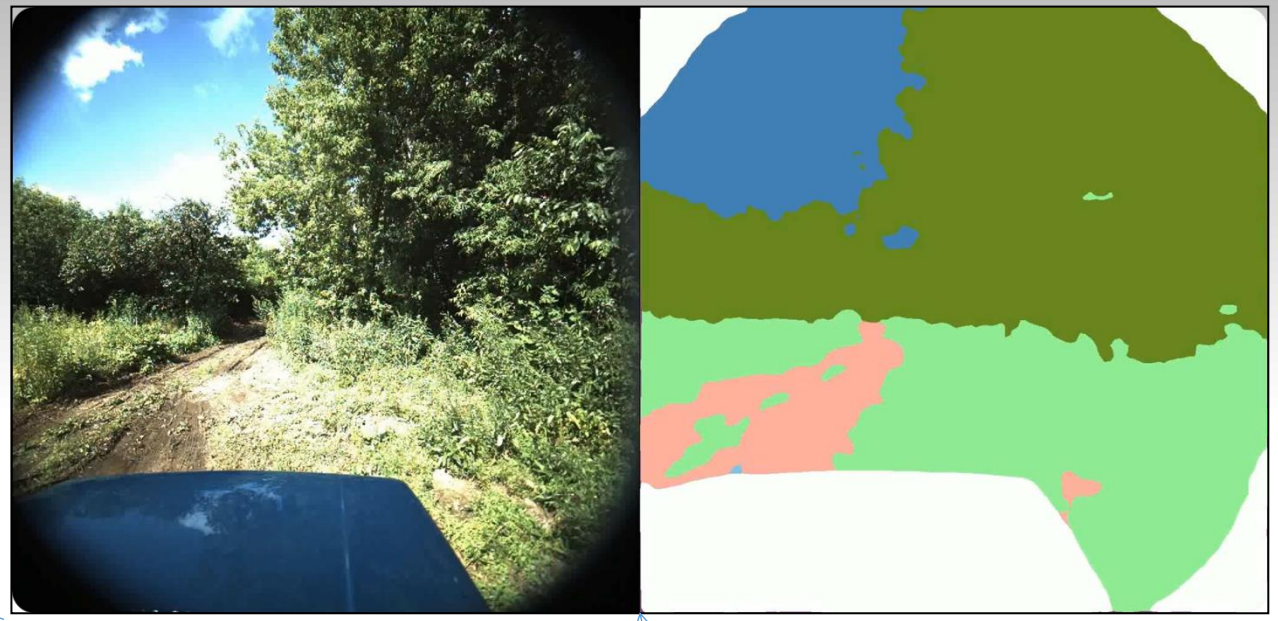
Собственная одометрия робототехнической платформы

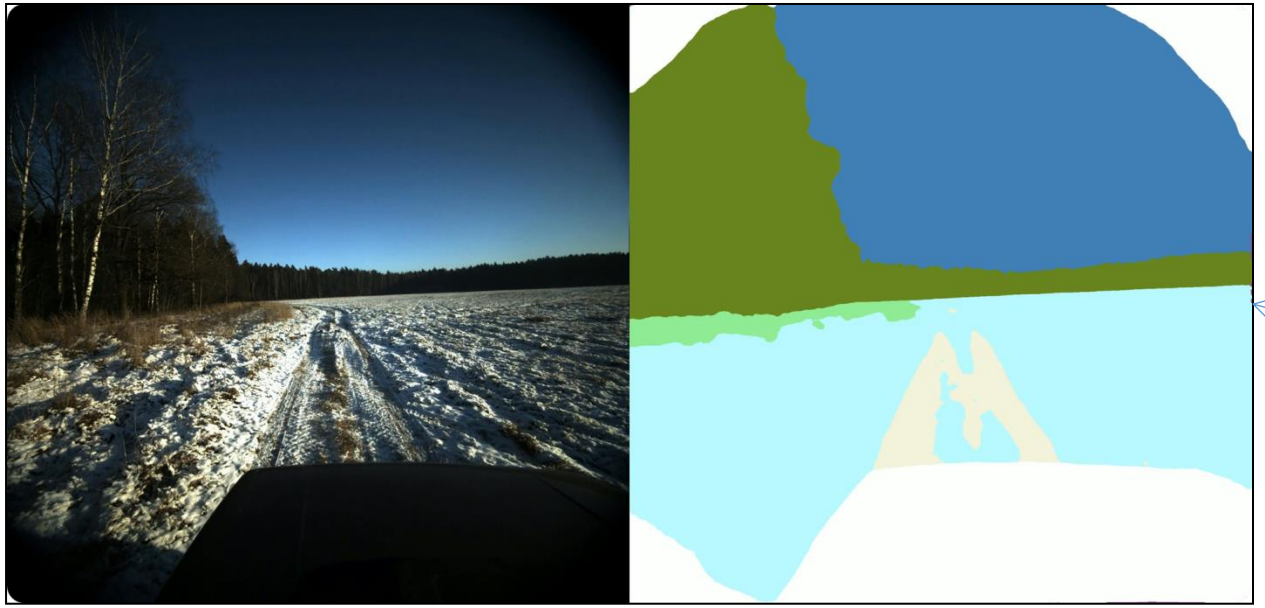


Трехмерное (возможно плотное) семантическое облако точек



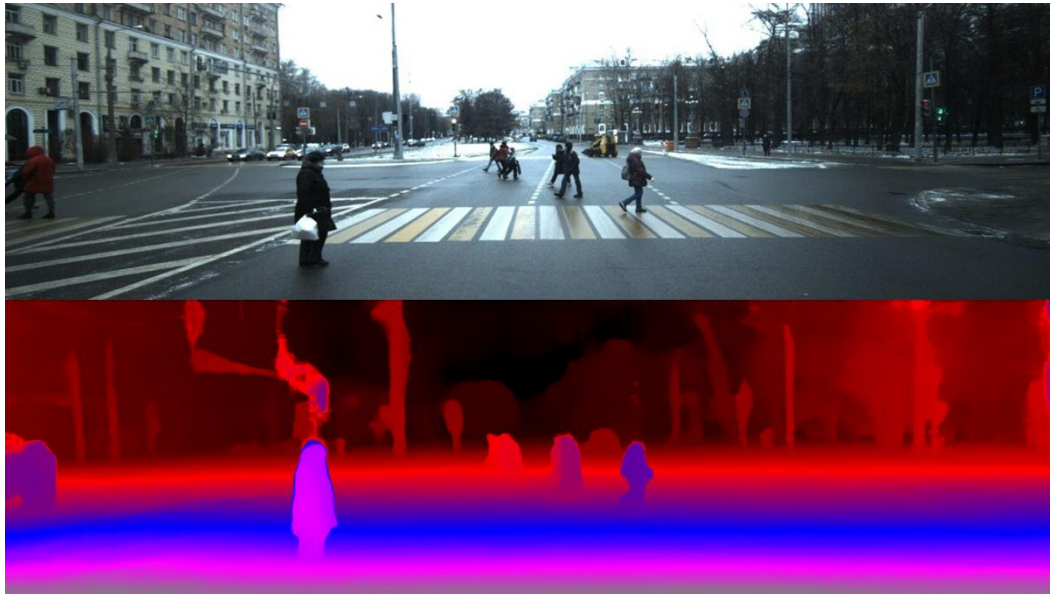
ТЕХНОЛОГИЯ СЕМАНТИЧЕСКОЙ МНОГОСПЕКТРАЛЬНОЙ СЕГМЕНТАЦИИ

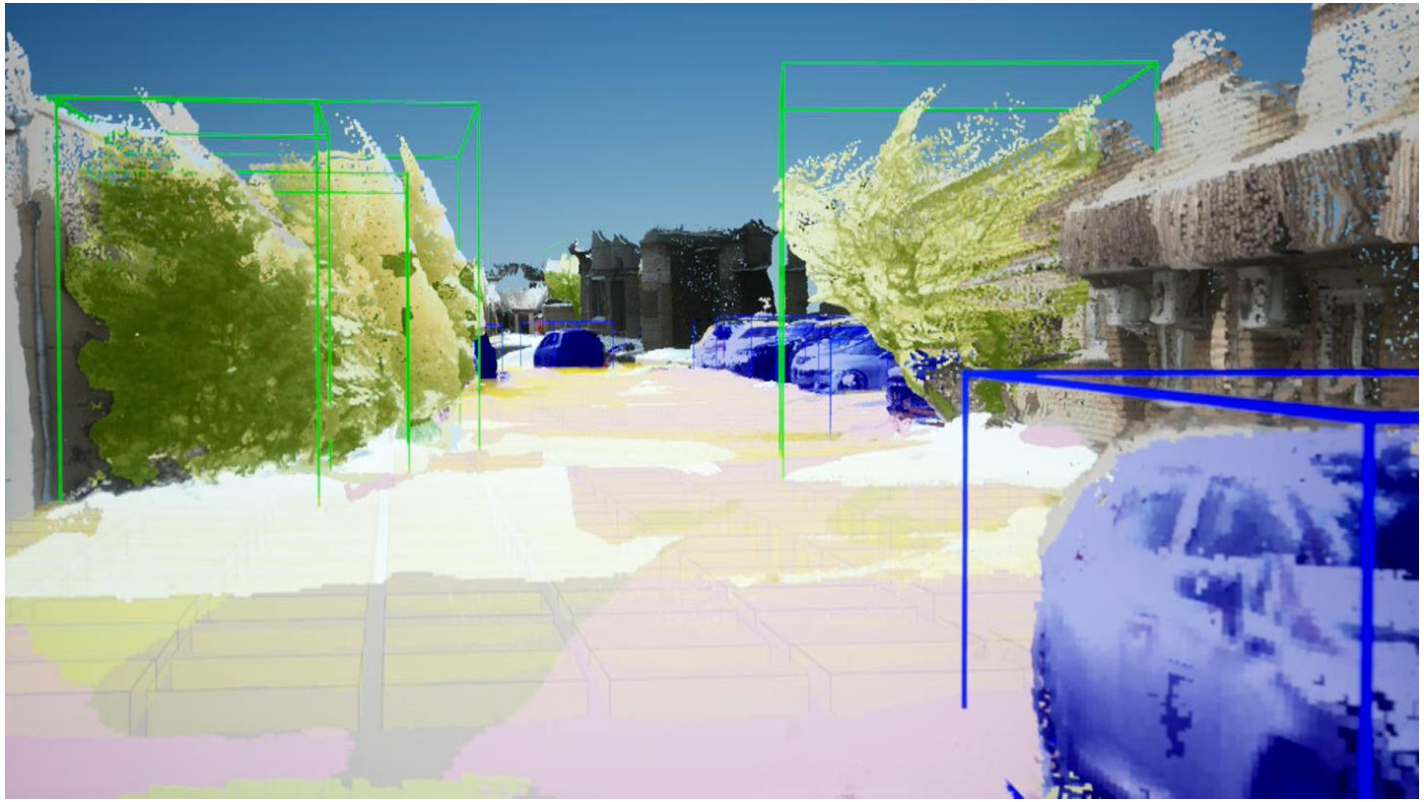




Семантическая сегментация многосенсорной мультиспектральной информации в различных условиях местности и сезона

ПЛОТНАЯ КАРТА ГЛУБИНЫ С ПОМОЩЬЮ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ





СЕМАНТИЧЕСКАЯ ТРЕХМЕРНАЯ МОДЕЛЬ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ

Сегодня пропустим

Интеллектуализация борта в гражданской авиации

Подр. Н.И.Сельвесюка, Подр.Е.Ю.Зыбина, Лаб. Д.В.Комарова, Подр. Б.В.Вишнякова

Задачи интеллектуализации ВС: борт ВС как источник многомодальных данных

Интеллектуальный борт = Автоматизация функций + Снижение нагрузки

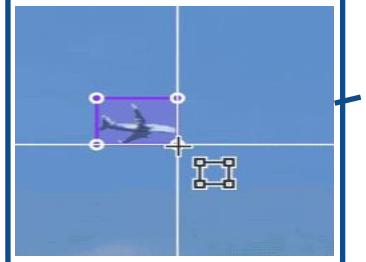
- Физиологическая составляющая
- Когнитивная составляющая

Прогноз > Предупреждение > Рекомендация > Действие

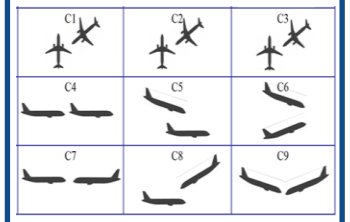
ЭКИПАЖ СРЕДНЕЙ КВАЛИФИКАЦИИ + ИНТЕЛЛЕКТУАЛЬНЫЙ КОМПЛЕКС БОРТОВОГО ОБОРУДОВАНИЯ = ЭКИПАЖ ВЫСШЕЙ КВАЛИФИКАЦИИ

Распознавание элементов ВПП и рулежных дорожек

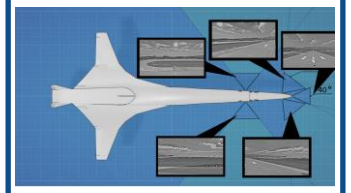
Распознавание объектов в воздухе и на поверхности



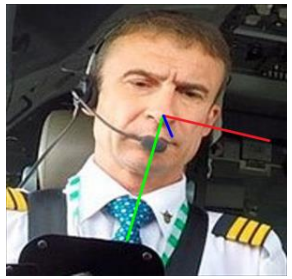
Определение конфликтных ситуаций при взлете и заходе на посадку



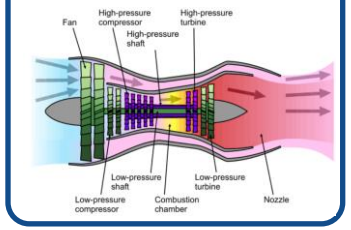
Формирование индикации закабинной обстановки в кабине без остекления



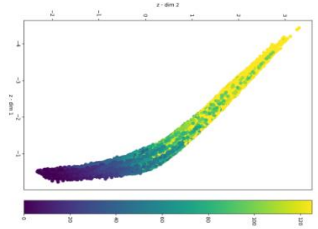
Мониторинга состояния и действий пилота



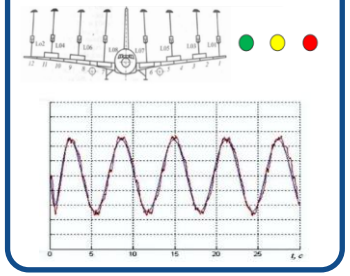
Обработка данных для мониторинга технического состояния ВС



Обработка массивов данных для функции предиктивной диагностики



Прогноз параметров полета на основе анализа текущих данных



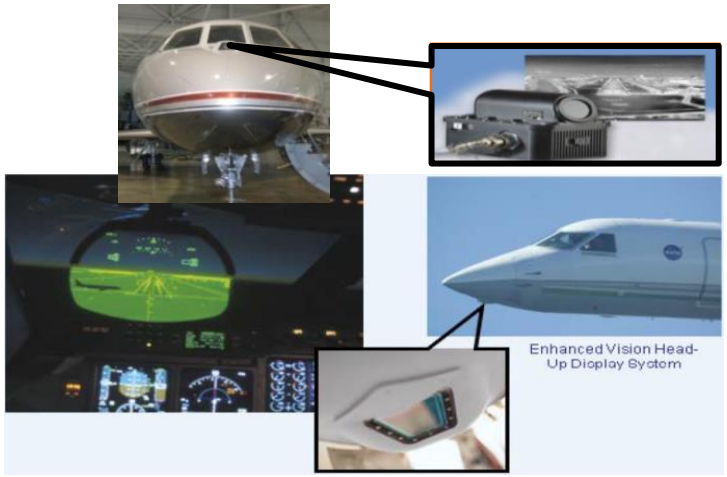
Виртуальные операционные помощники пилота

Виртуальный второй пилот

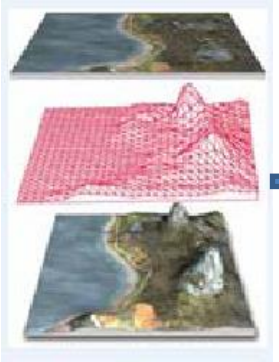
Виртуальный пилот

Интеллектуальные технологии для авиационных систем улучшенного, синтезированного и комбинированного видения

Система улучшенного видения (СУВ)



Система синтезированного видения (ССВ)



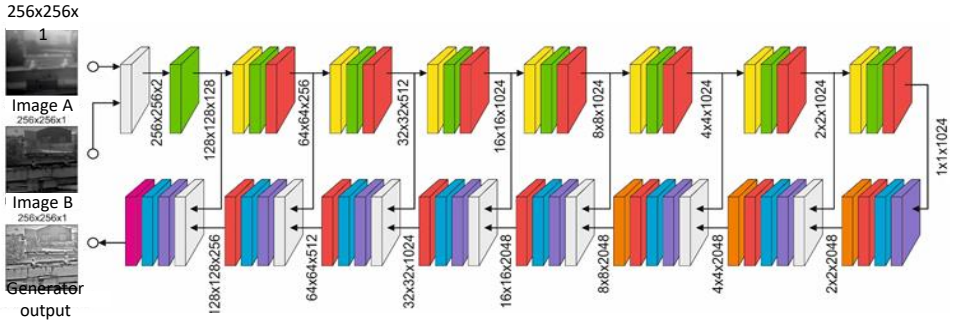
Система комбинированного видения (СКВ)

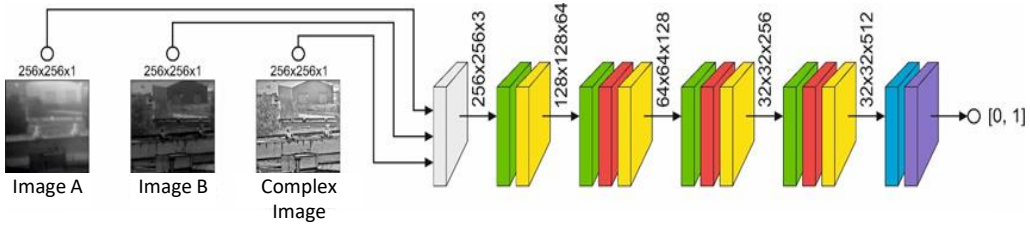
СУВ – источник многоспектральных данных

ССВ – добавляет другие модальности

Комплексирование многоспектральных изображений СУВ

Задача СУСКВ:
повышение ситуационной осведомленности экипажа
при полете, посадке и выполнении задач в ограниченных условиях видимости



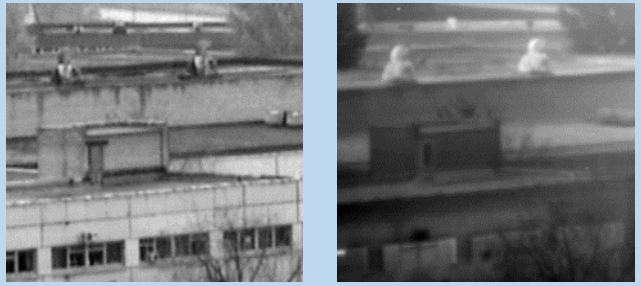


Архитектура ГНС и результат комплексирования

ТВ ИК

Laplacian pyramid FusionNet

Удаление артефактов



ТВ ИК

Laplacian pyramid FusionNet

Повышение детальности

ТВ ИК

Laplacian pyramid FusionNet

Автоматическая привязка

Семантическая сегментация для СУСКВ, динамическая привязка СУВ к ССВ

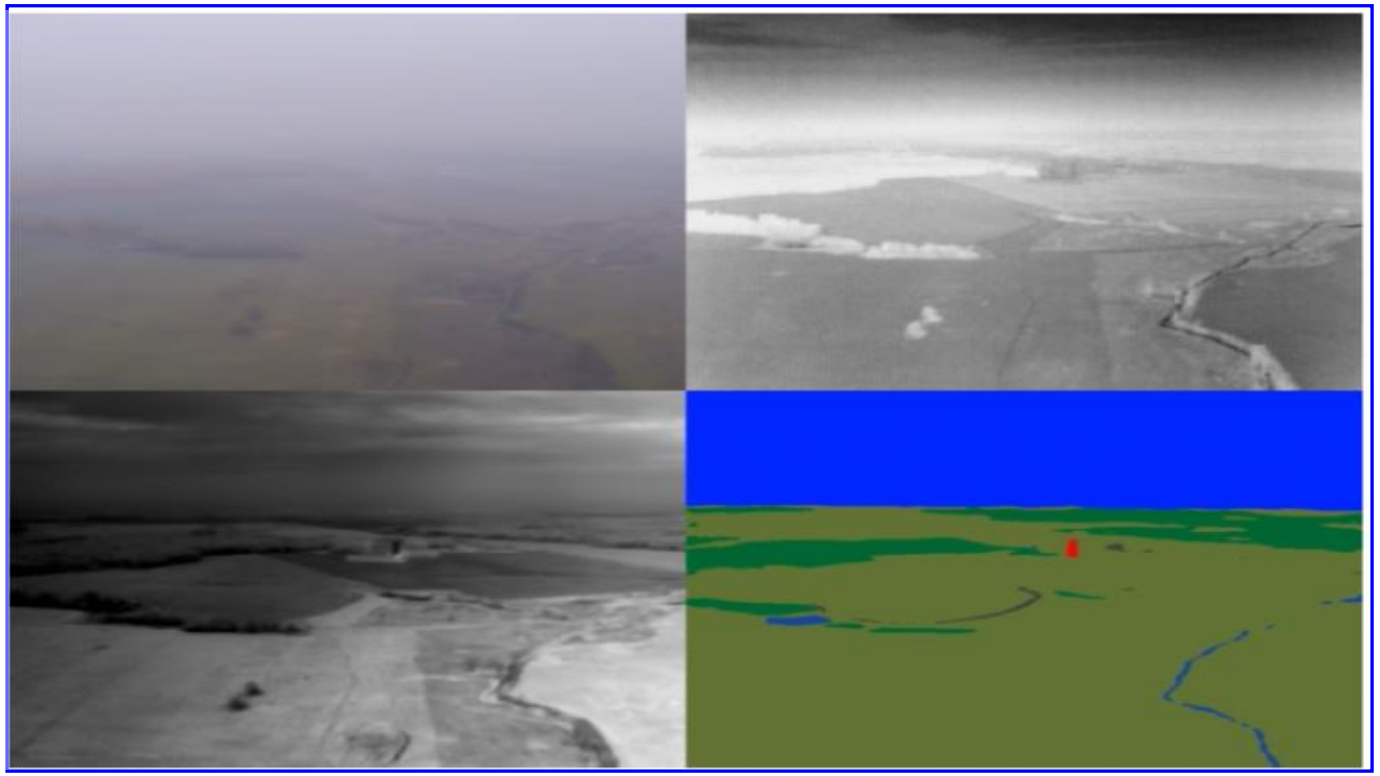
Задача СУСКВ:
повышение ситуационной осведомленности экипажа
при полете, посадке и выполнении задач в ограниченных условиях видимости

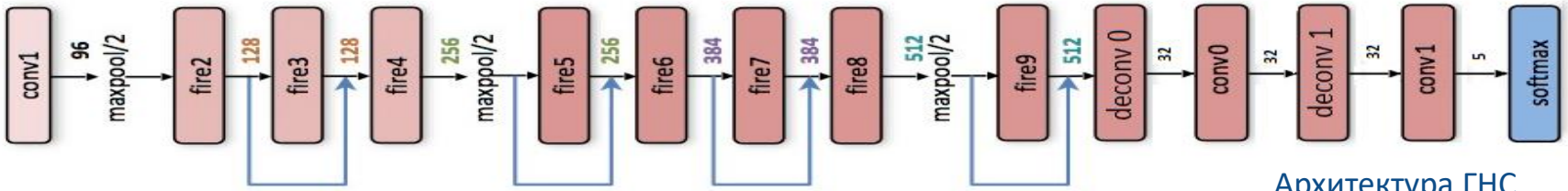
Преимущества:

- Возможность реализации режима комбинированного видения с использованием **дополненной реальности**

- Автоматическое обнаружение и отображение **некартографированных объектов**

- Помощь летчику в **интерпретации многоспектральных данных**



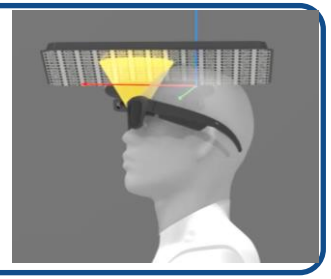


Архитектура ГНС

Мониторинг состояния пилота и человеко-машинный интерфейс

Определение положения координат глаз и направления взгляда пилота на основе алгоритмов машинного обучения с применением нескольких камер

Определение положения и пространственной ориентации головы пилота на основе нейронных сетей глубокого обучения



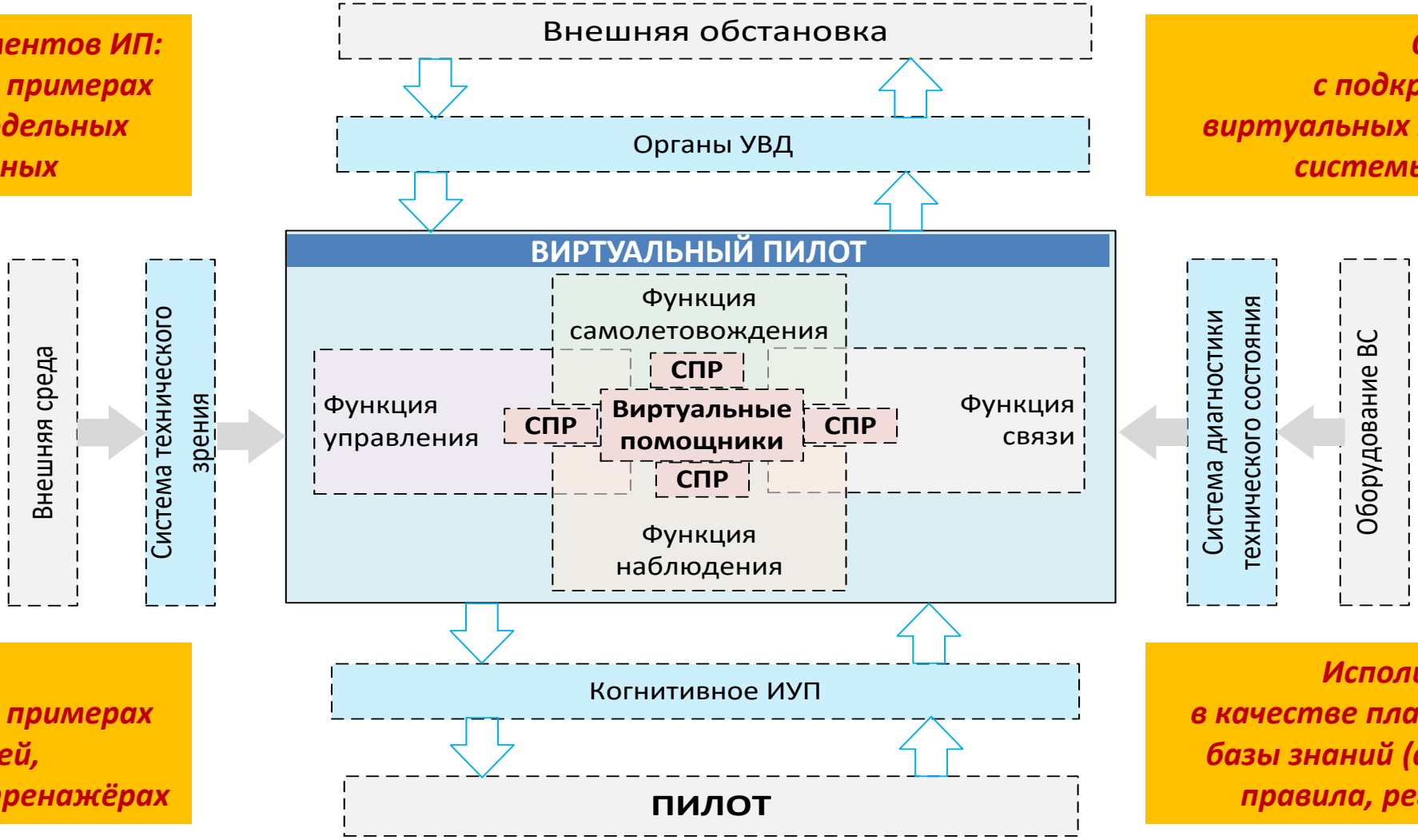
Распознавание речи, обработка естественного языка, интеллектуальный голосовой помощник для обработки голосовых команд экипажа

Источники многомодальных данных

Виртуальный пилот и виртуальные помощники: технологии ИИ

Обучение элементов ИП: с учителем на примерах реальных и модельных сенсорных данных

Обучение ИП: с подкреплением на виртуальных моделях ВС и системы управления



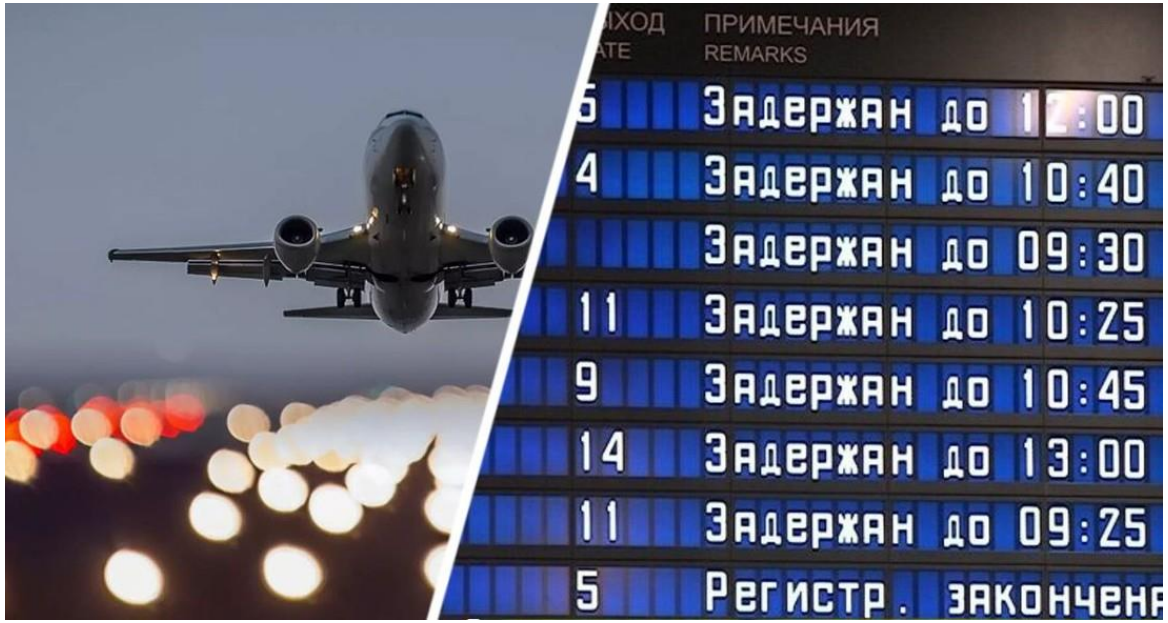
Обучение ИП: с учителем на примерах действий людей, собранных в тренажёрах

Использование LLM в качестве планировщика и базы знаний (авиационные правила, регламенты,..)

Анализ многомодальной информации для выработки управления

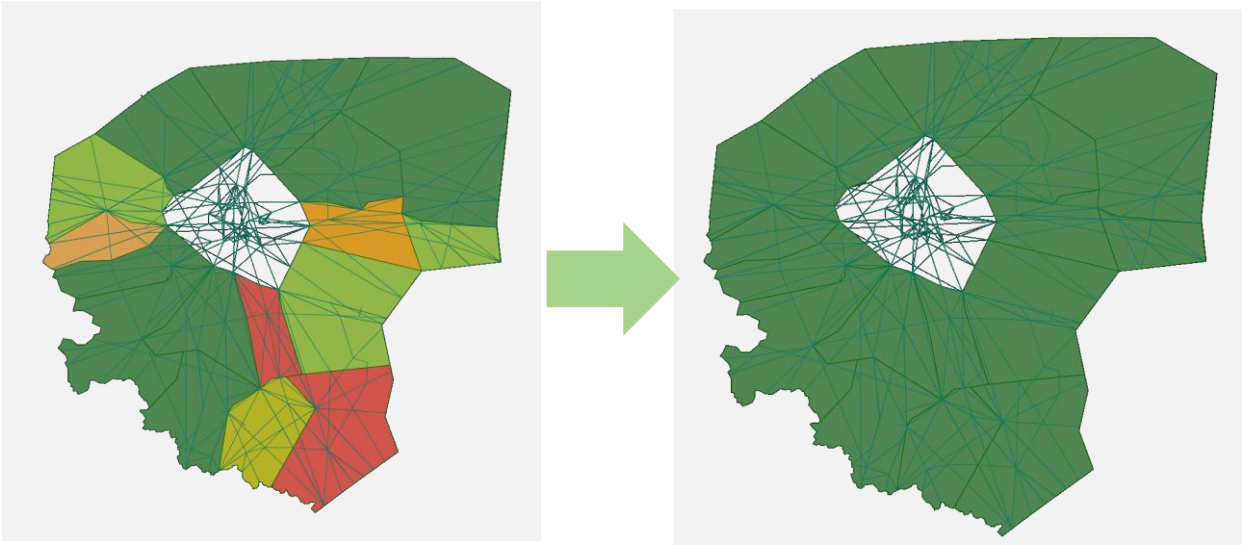
Лаб. Н.В.Трухтанова, Подр. Л.В.Вишняковой, Лаб. И.В.Сгибнева

Обучение с подкреплением для принятия решений при формировании потока воздушного движения



Интеллектуальные агенты должны обеспечивать формирование рациональных потоков воздушного движения в части назначения времени вылета воздушных судов
с учетом:

- минимизации суммарного времени задержек воздушных судов
- отсутствия превышений нормативов пропускной способности секторов управления воздушным движением
- глубины прогнозирования потоков воздушного движения не менее 2 часов

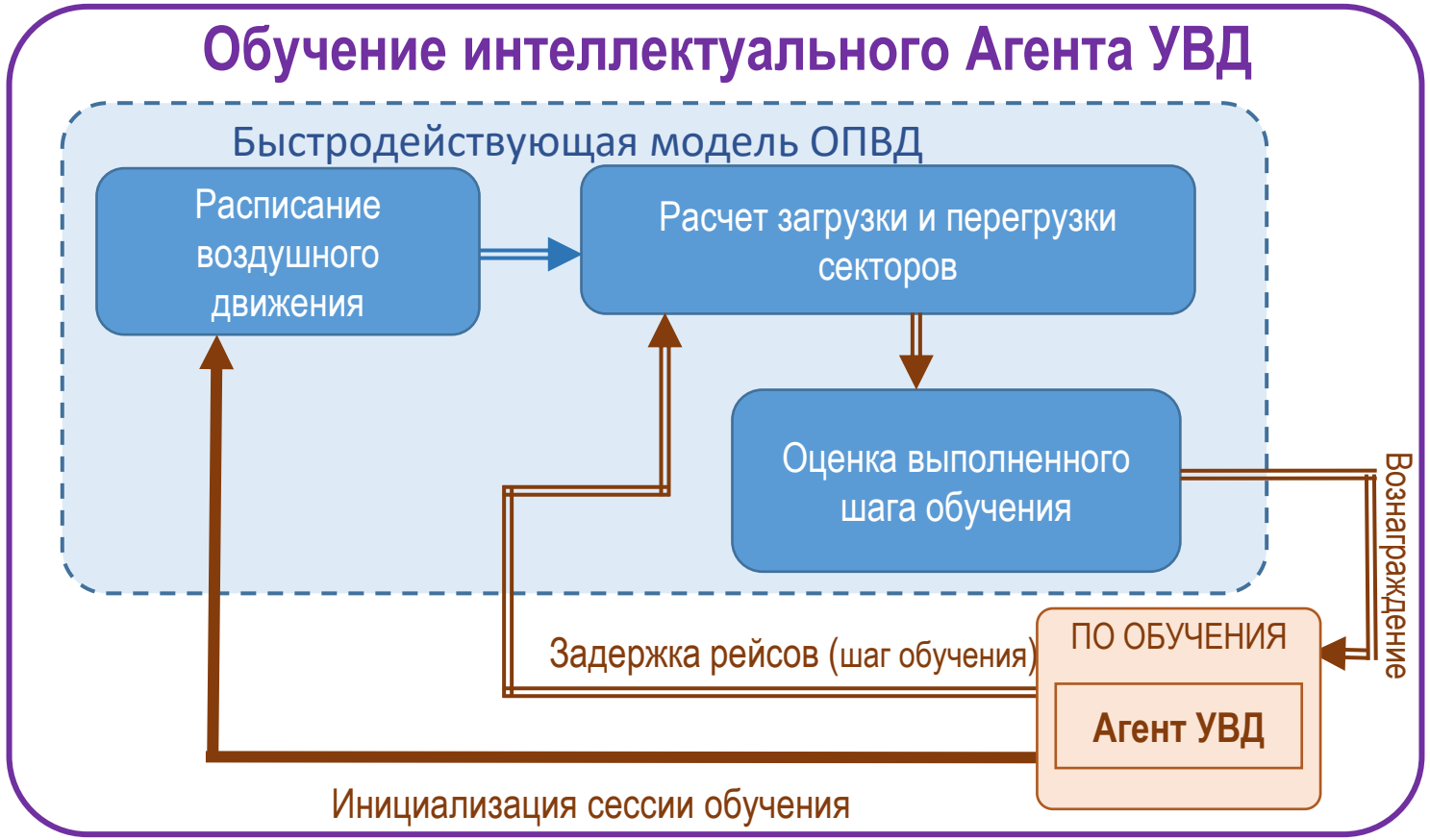


Классические методы формирования слотов вылета

Недостатки классические методов:

- неоптимальное решение
- при высокой интенсивности ВД поиск решений требует более сложных алгоритмов

Применение RL для задачи слотирования



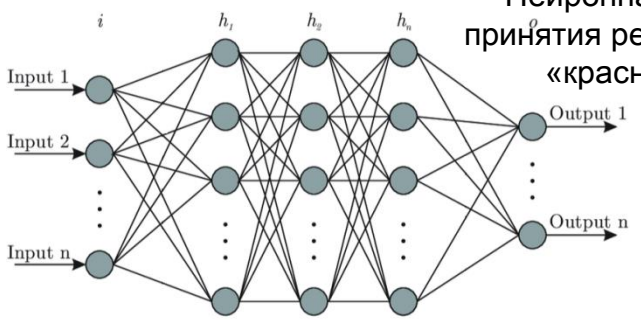
Использование обученного интеллектуального агент УВД позволит решить задачу с назначение меньших задержек рейсам

Процесс обучения агентов

Агент

Входные данные

- Электронная карта местности
- Защищаемые объекты
- Силы обороны



Нейронная сеть принятия решений за «красных»

Действие

Варианты решений по обороне

Среда

Вознаграждение

Численная интерпретация результативности отражения атаки

Агент

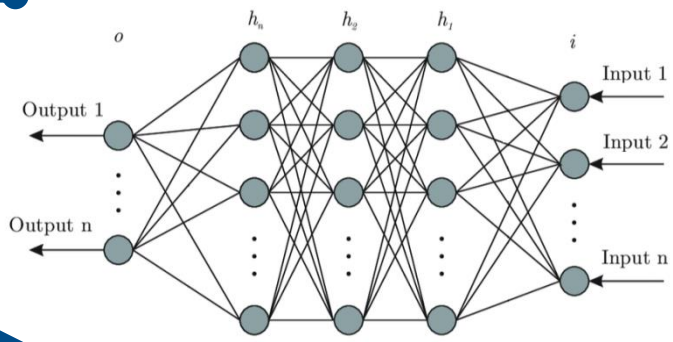
Действие

Варианты решений по нападению

Среда

Вознаграждение

Численная интерпретация результативности атаки



Нейронная сеть принятия решений за «синих»

Входные данные

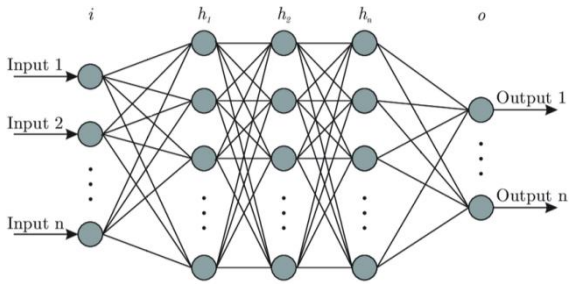
- Электронная карта
- Защищаемые объекты
- Силы нападения

Обучение на основе антагонистической игры

Действие

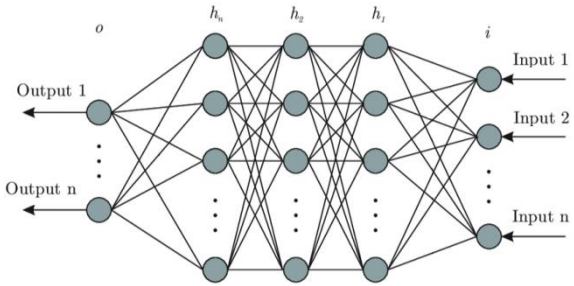
Действие

Агент



Имитационная модель (среда моделирования)

Агент



Вознаграждение

Вознаграждение

Результативность обороны

Результативность атаки

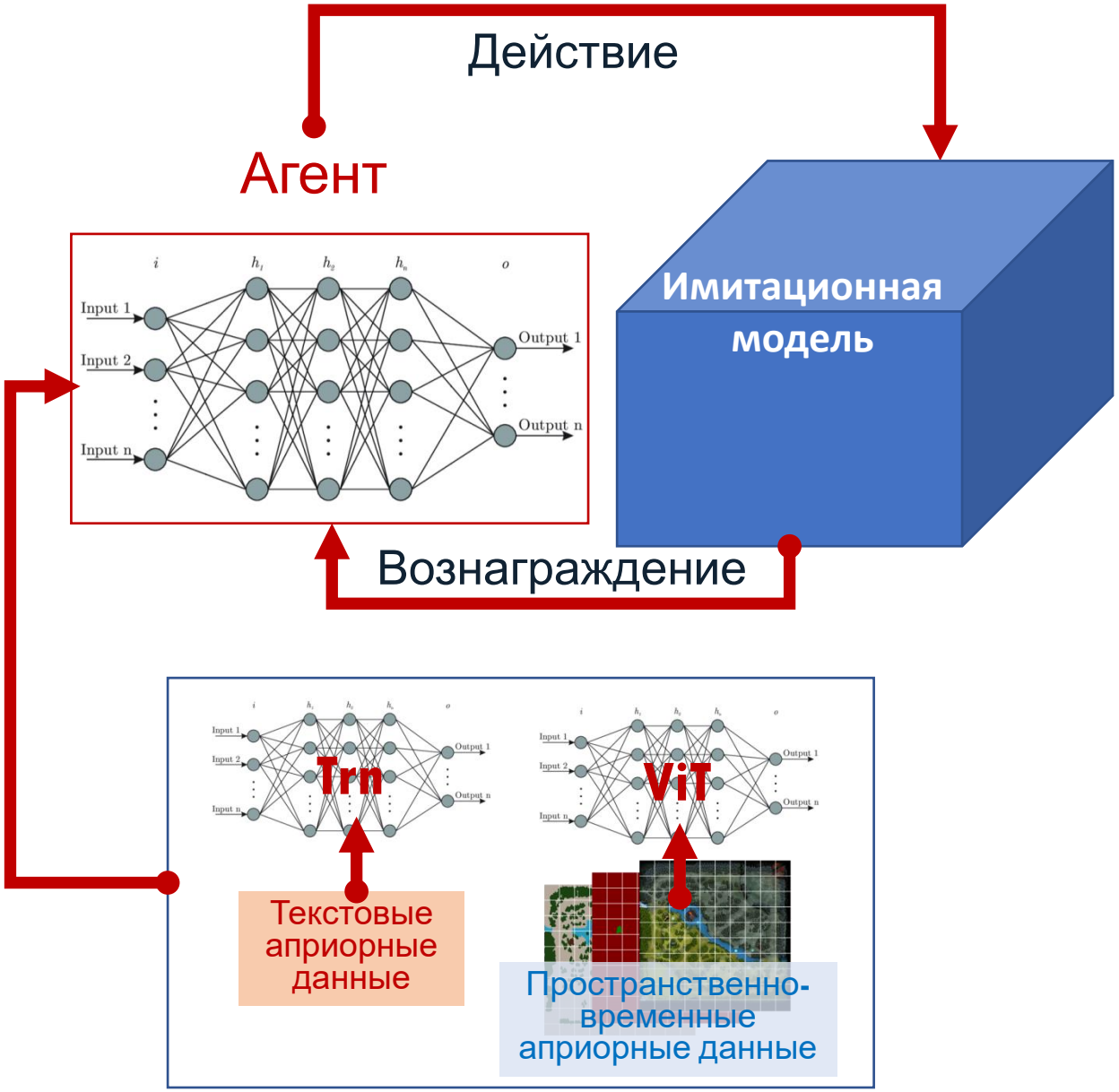
Входные данные

Входные данные

Оптимизация методом обучения с подкреплением:

- Антагонистическая игра
- Статичная карта (один район)
- Количество имитаций 192 000
- Время оптимизации 30 минут

Обучение без использования модели мира



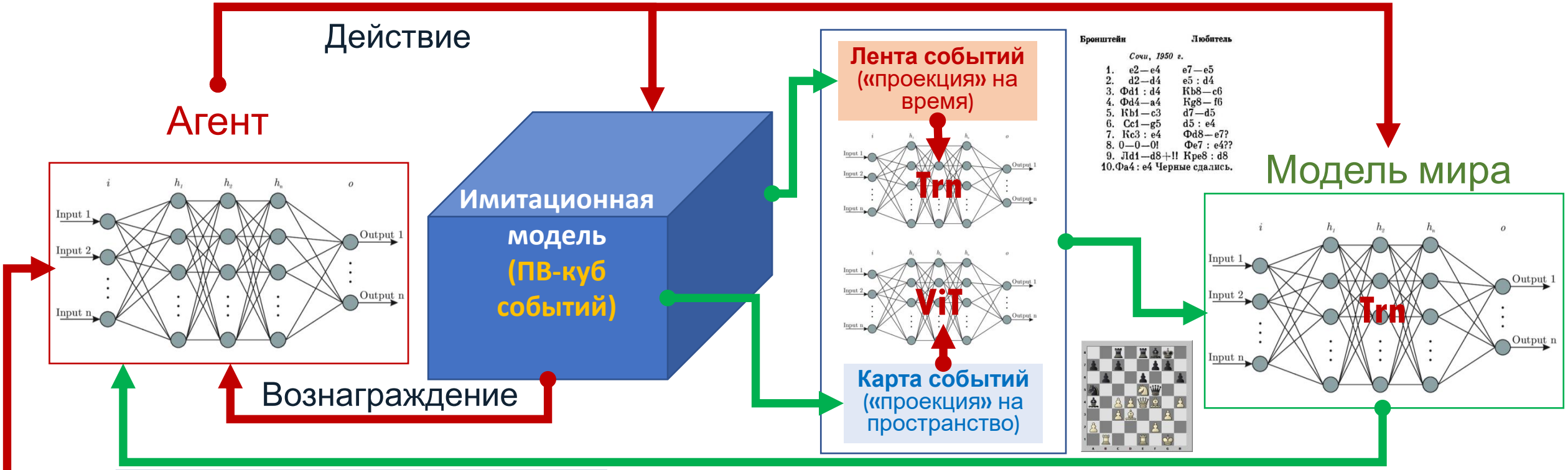
Проблемы:
- отложенное вознаграждение
- сложность оценки влияния каждого принятого решения на финальный результат (вознаграждение)

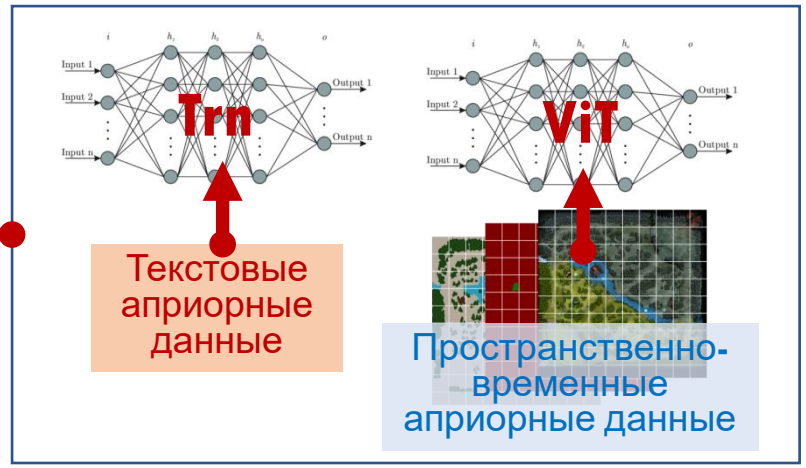
Возможный путь решения: добавить дифференцируемую модель мира, которая может использоваться для обучения с обратным распространением ошибки в дополнение к RL

Сложность: модель мира нужно специально изобретать под каждую конкретную задачу RL для каждой среды моделирования.

Предложенный подход:
- использовать универсальную архитектуру событийной модели мира, которая обучается под разные задачи и разные среды моделирования
-использовать бимодальное кодирование информации о событиях

Обучение с использованием «событийной» модели мира





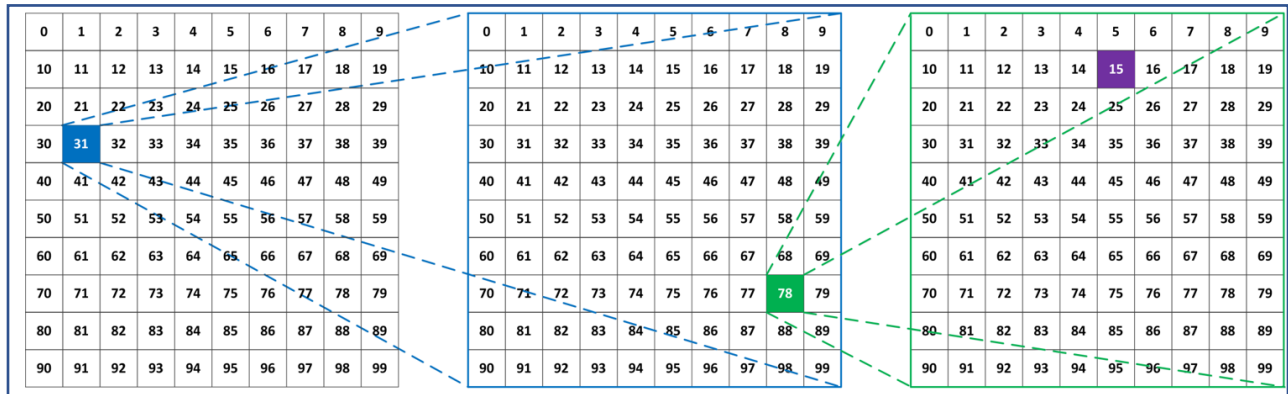
Предсказание вознаграждения

Особенности подхода:

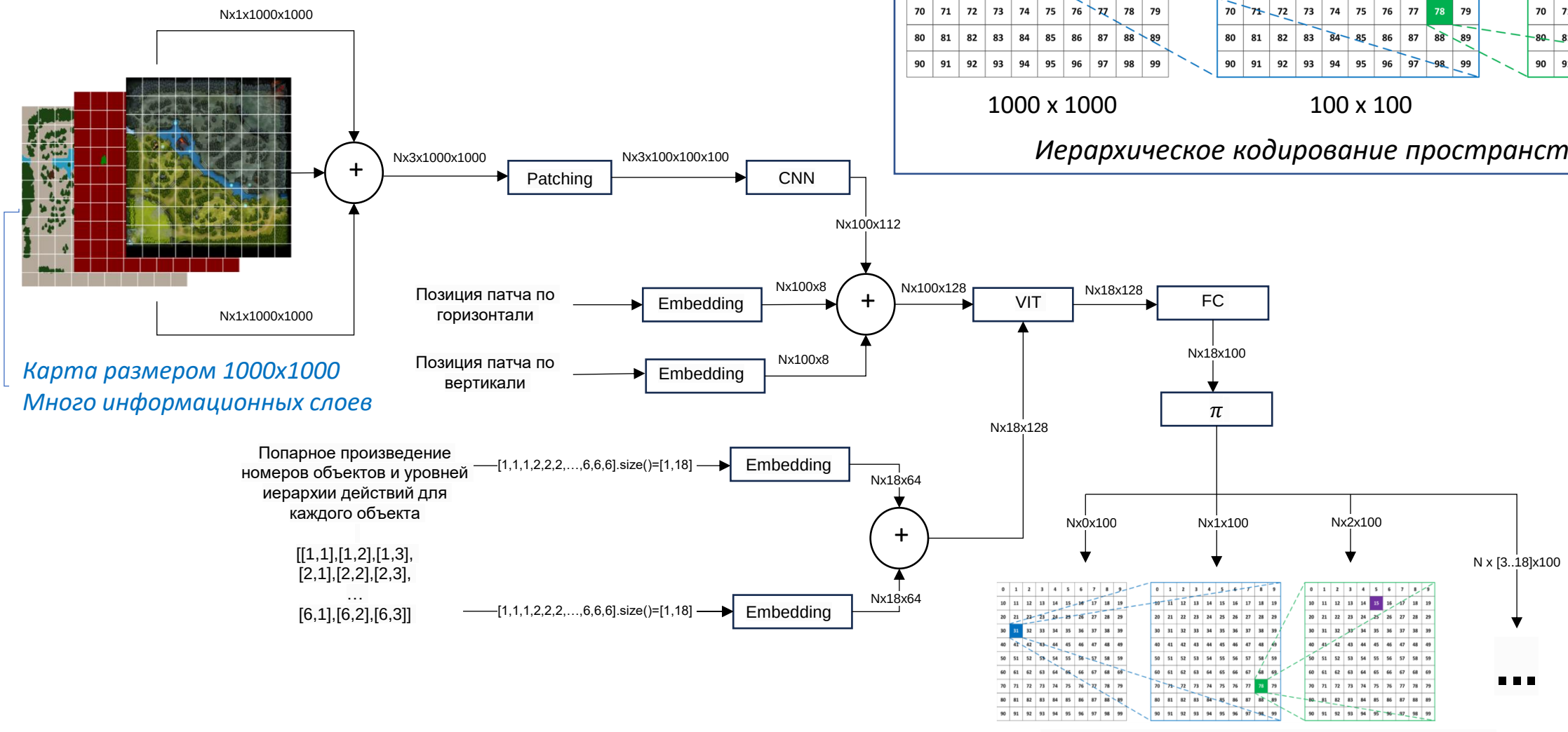
- Модель мира – мультимодальный трансформер
- События = **(время, место, тип, затронутые объекты, задействованные акторы)**
- Двойное кодирование – лента событий и карта событий
- Модель мира обучается предсказывать вознаграждение
- Предсказание вознаграждения – дифференцируемая функция потерь

Пример обучения с подкреплением для принятия решений в тактической игре «красные» против «синих»

Структура нейронной сети агента защиты



Иерархическое кодирование пространства действий

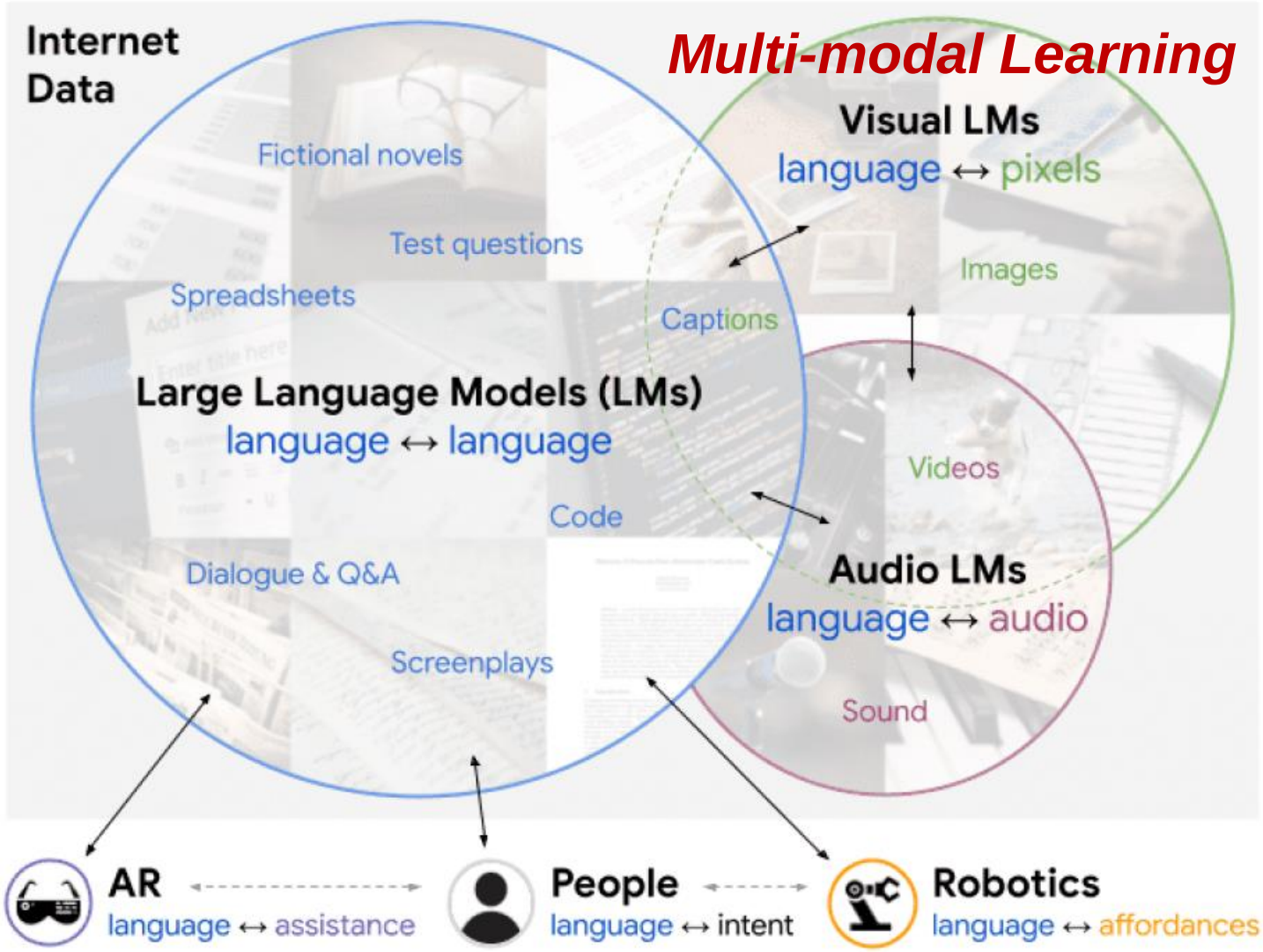


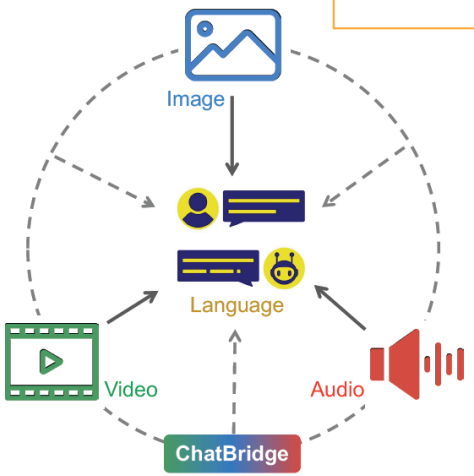
Карта размером 1000x1000
Много информационных слоев

18 действий (по 3 уровня иерархии для каждого объекта)

Socratic Models: решение мультимодальных задач коллективом агентов. От GhatGPT к чату LLM-агентов

Возможна ли универсальная схема ИИ для всех задач?





Гипотеза: для этого все входные и выходные данные нужно перевести в текстовую модальность, с которой работают LLM.

Проблема: для обучения нужна огромная мультимодальная база

Решение: научить модели использовать запросы, и пусть модели общаются между собой

«Сократические модели» это способ построения расширяемого коллектива ИИ-агентов с «открытым» списком задач и типов данных, для работы с которыми добавляются новые агенты

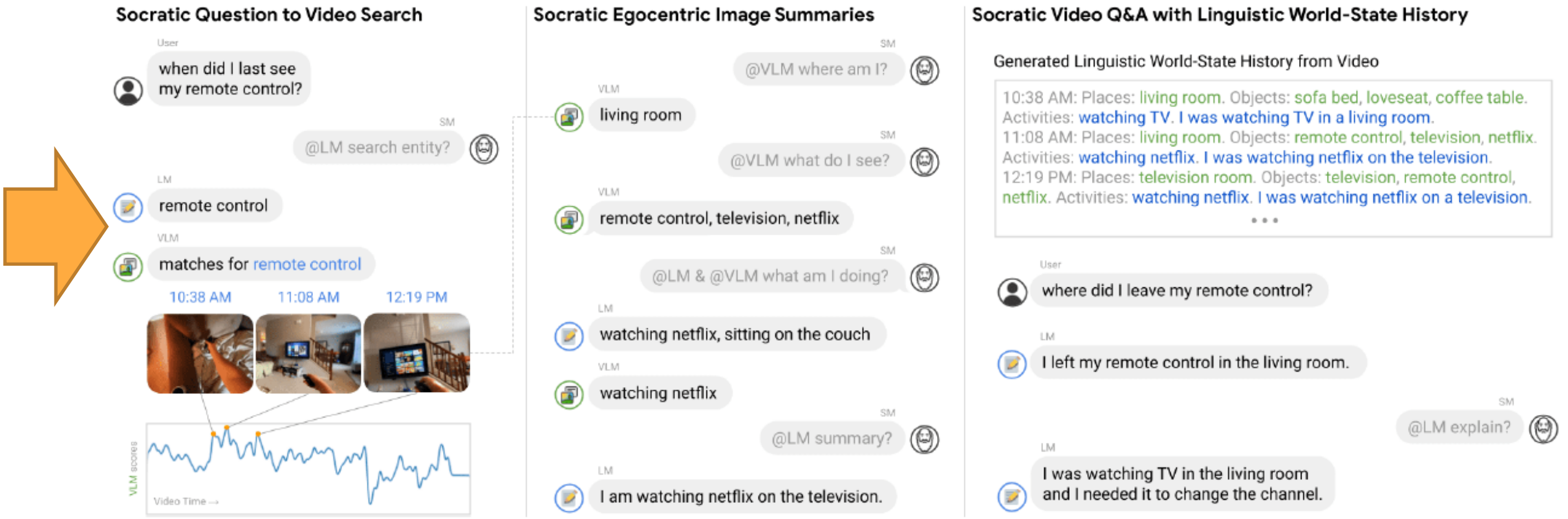
Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Google, 2022

Socratic Models: решение мультимодальных задач коллективом агентов. От GhatGPT к чату LLM-агентов

Возможна ли универсальная схема ИИ для всех задач?

Метод Сократа – решение задачи путем диалога

Внутри работа «сократической модели» действительно выглядит как чат ИИ-агентов!

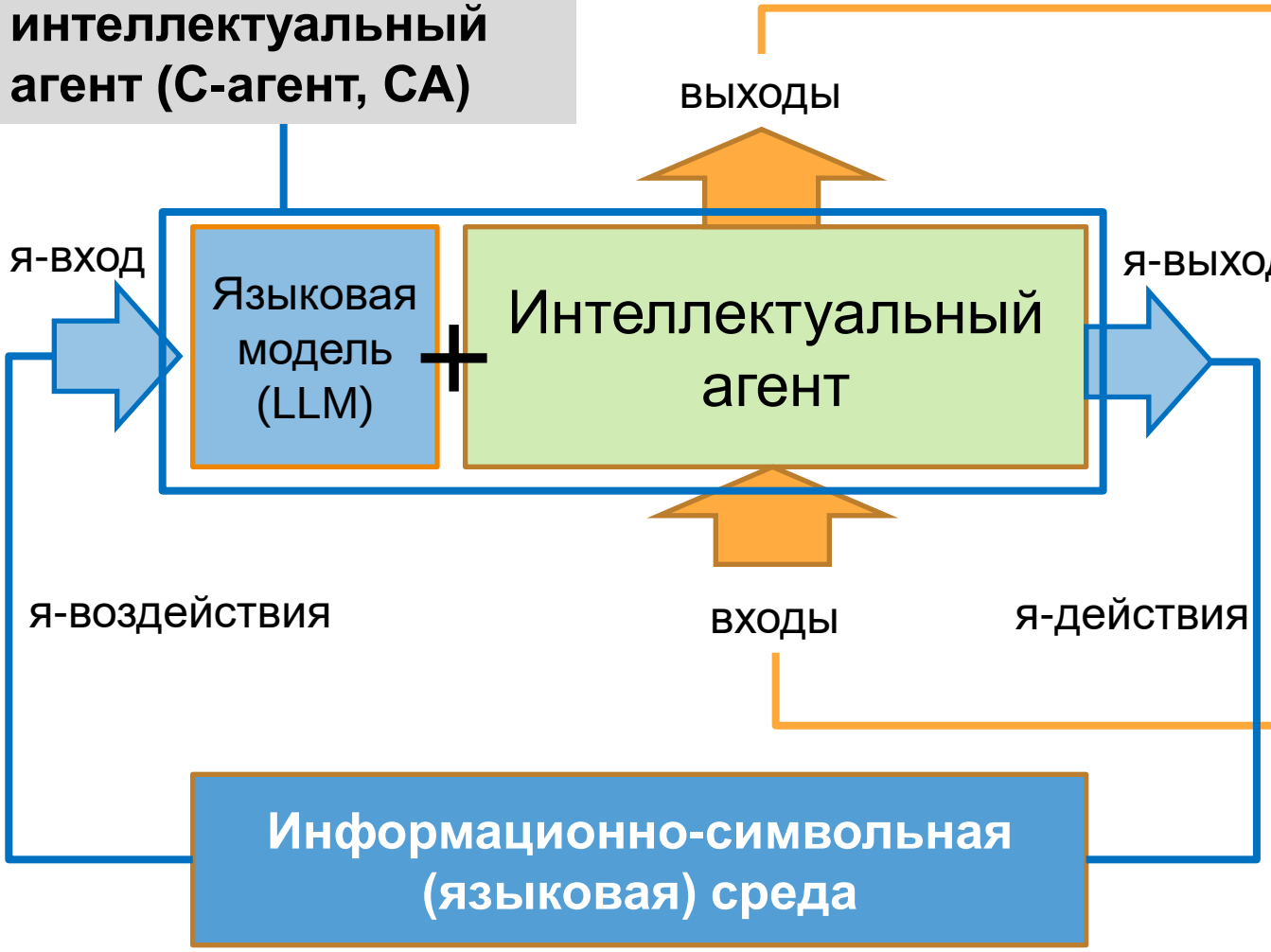


from LMs are blue, VLMs green, ALMs purple, prompt text gray, user inputs magenta, VLM-chosen LM outputs green-underlined blue, and ALM-chosen LM outputs purple-underlined blue.

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Google, 2022

Обобщение: сократическая схема универсального интеллекта

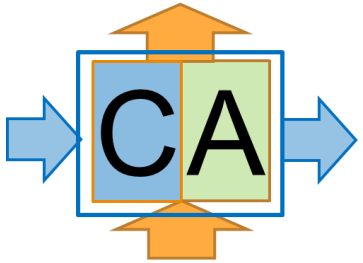
Сократический интеллектуальный агент (С-агент, СА)



Работа и обучение в языковой (символьной, культурной, социальной) среде

Физическая среда (симулятор)

Работа и обучение в физическом мире

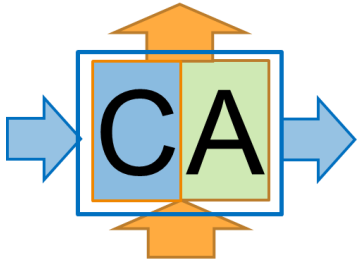


С-агент живет одновременно в двух реальностях - физической и языковой (символьной, культурной) и выучивает 2 типа поведения - физическое и языковое, которые вместе помогают ему быть эффективным одновременно в обеих средах

Обобщение: сократическая схема универсального интеллекта

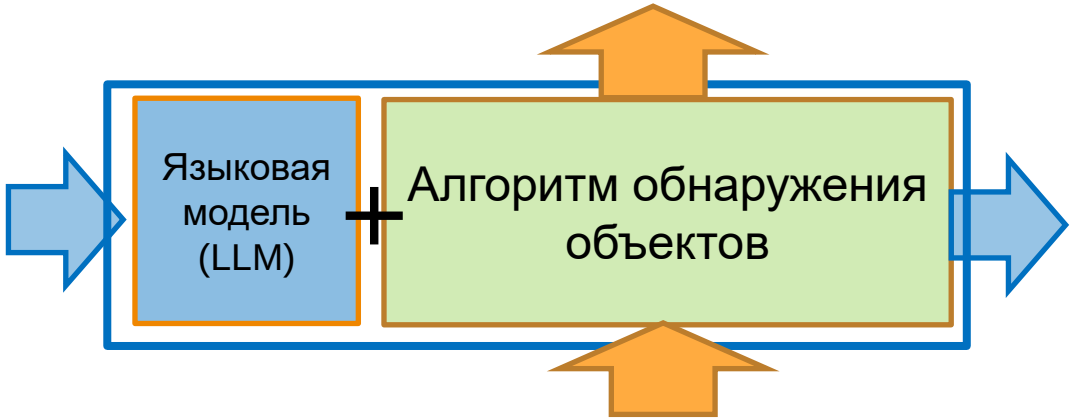
Уровни СА

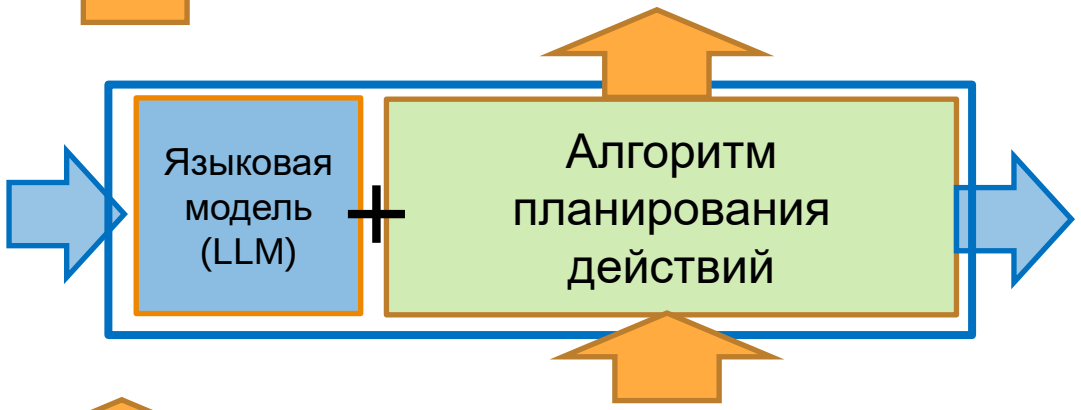
1. Функциональный (слабый) СА

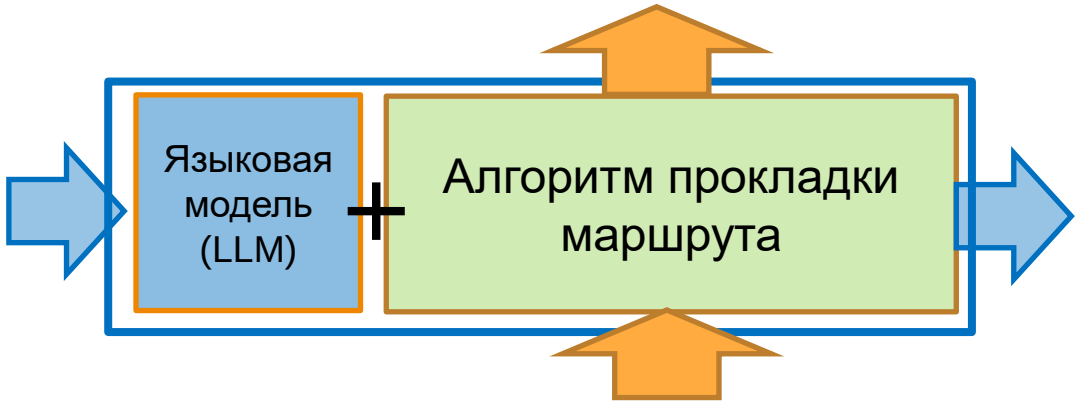


Простейшим «слабым» СА может быть любой алгоритм анализа данных или управления, обёрнутый языковым вводом-выводом

Наша расширенная интерпретация «бимодальных агентов»:
одна модальность язык, вторая «модальность» - частная функция анализа данных или синтеза данных или выработки управления, планирования, принятия решения



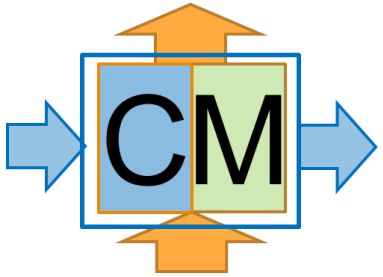




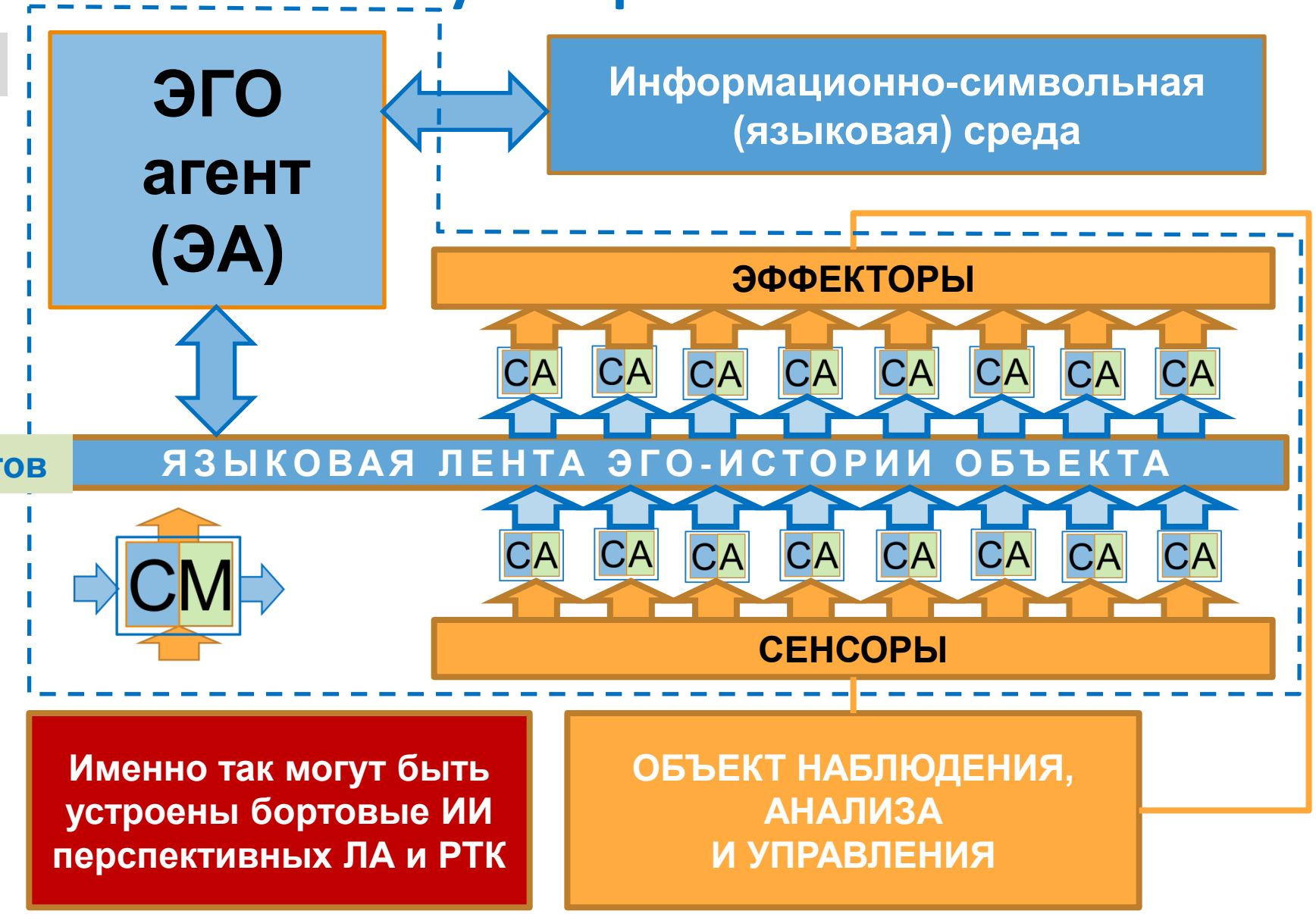
Обобщение: сократическая схема универсального интеллекта

Уровни СА

1. **Функциональный (слабый) СА**
2. **Универсальная («сильная») эгоцентрическая сократическая модель (С-модель, СМ)**



***Сократическая модель** это **коллектив слабых с-агентов**, связанных с некоторым **физическим объектом** или технической системой и **единой лентой упорядоченных событий** (**чатом**) = **«личной» эгоцентрической истории** (+ **эго-агент!**)*



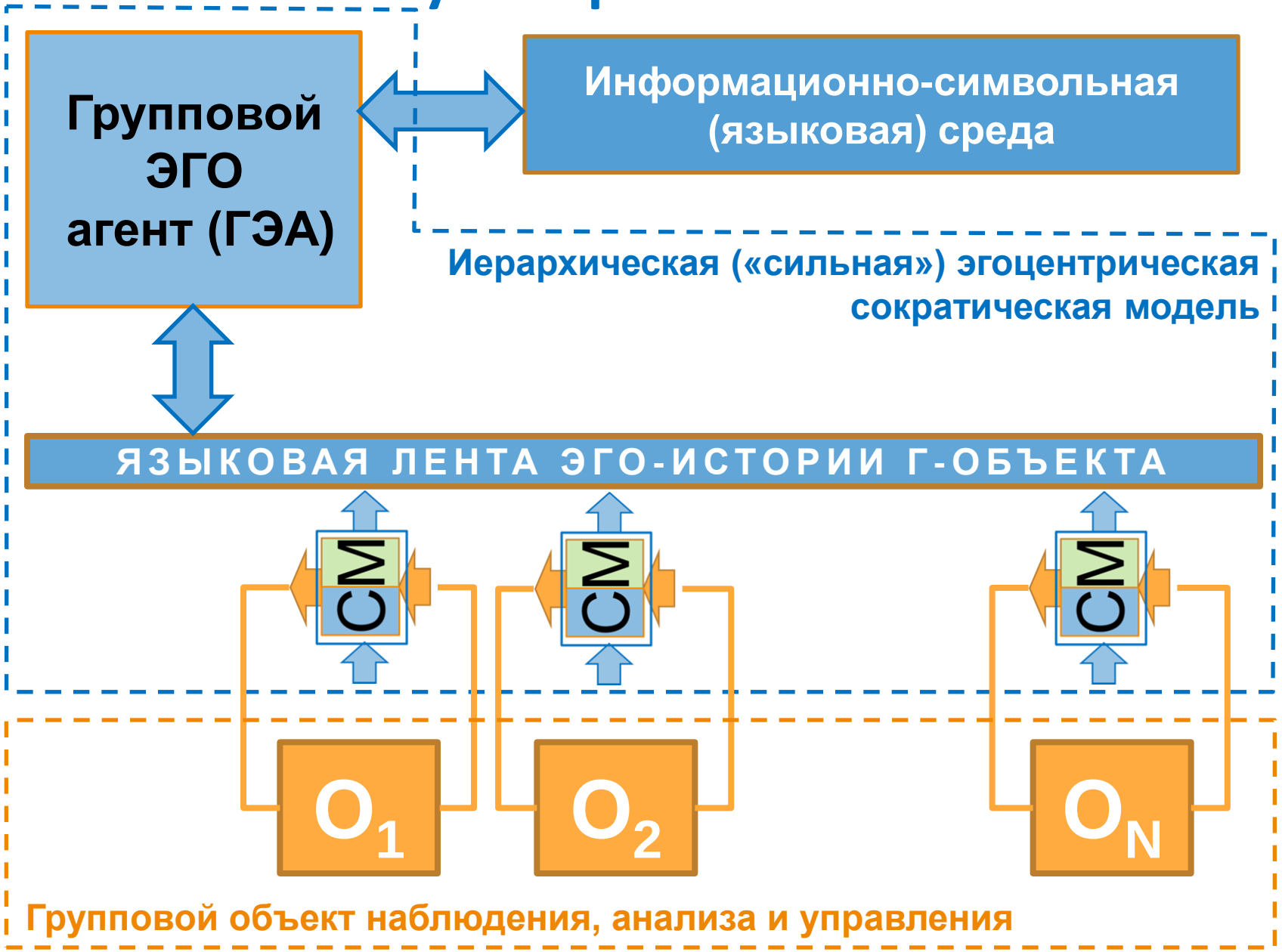
Обобщение: сократическая схема универсального интеллекта

Уровни СА

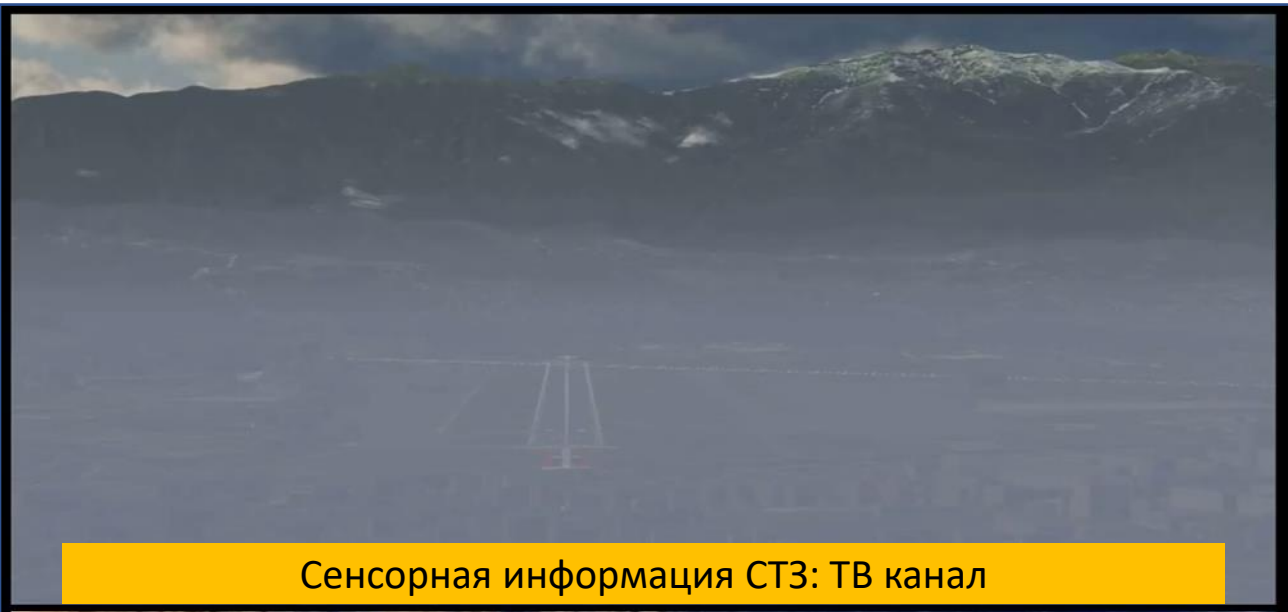
1. Функциональный (слабый) СА
2. Универсальная («сильная») эгоцентрическая сократическая модель (С-модель, СМ)
3. Иерархическая («сильная») эгоцентрическая сократическая модель (ИС-модель, ИСМ).

Любое число уровней иерархии

ИСМ** это **коллектив сильных С-моделей**, связанных с некоторым **групповым физическим объектом** или технической системой и **единой лентой** (**чатом**) **упорядоченных событий «групповой» эго- истории (+ групповой эго-агент!)

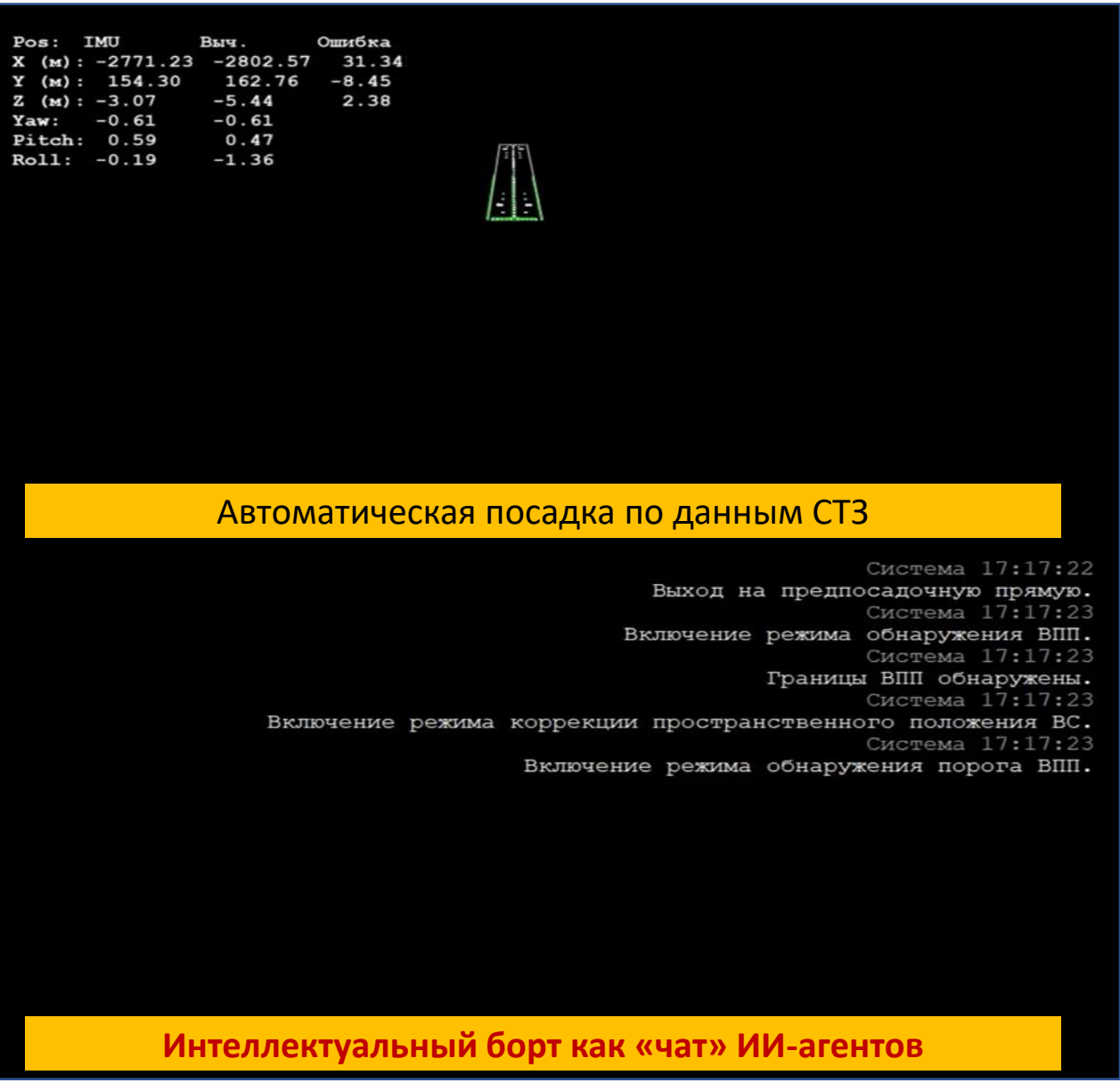


Моделирование сократического ИИ на этапе посадки ВС ГА



Сенсорная информация СТЗ: ТВ канал

Система улучшенного видения ТВ канала



Автоматическая посадка по данным СТЗ

Интеллектуальный борт как «чат» ИИ-агентов

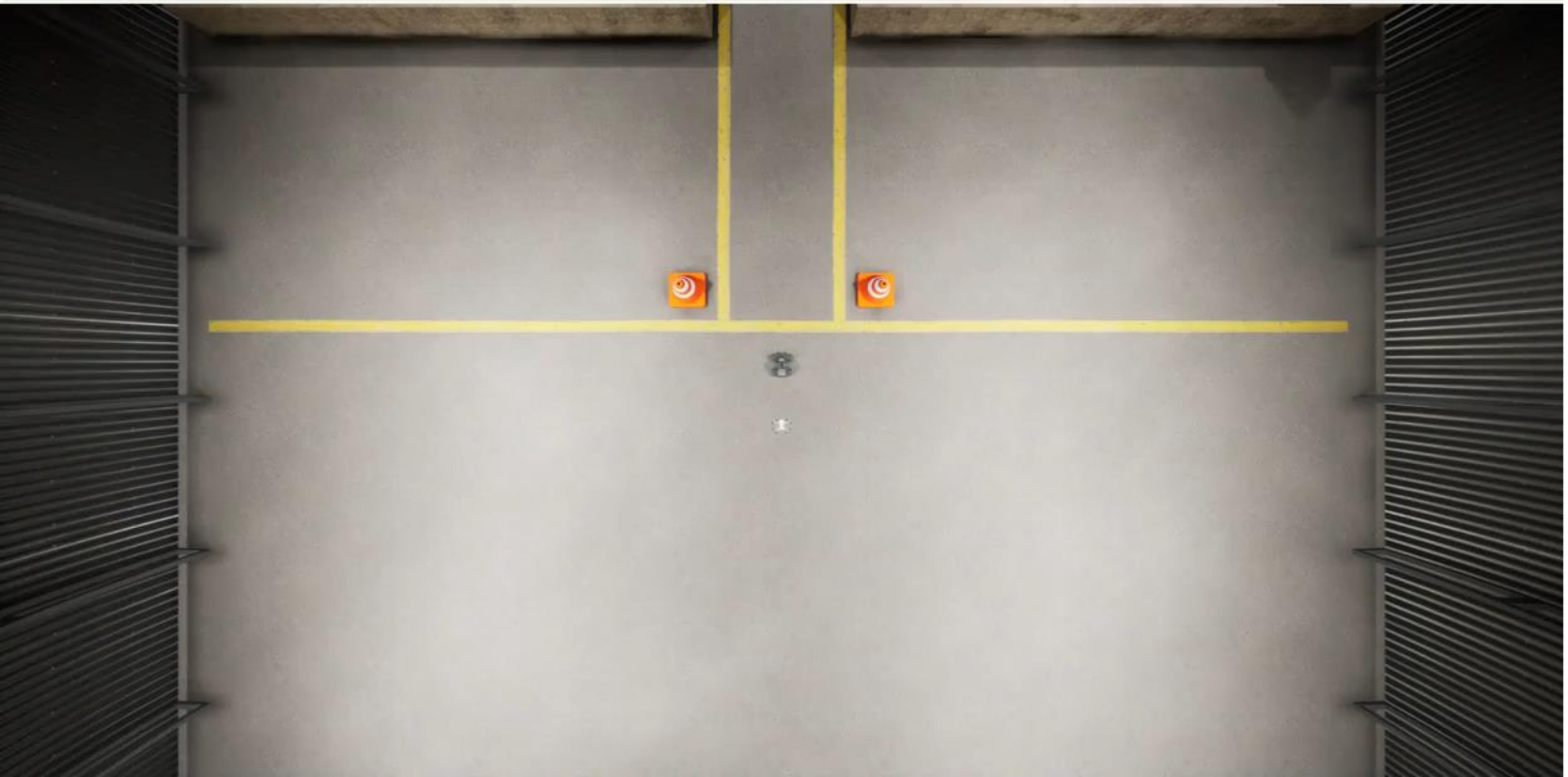
Моделирование сократического ИИ для VLA-задачи группы РТК

Агент семантической сегментации

Агент построения карты глубины

Чат интеллектуальных агентов

Обзорная камера



Введите сообщение...

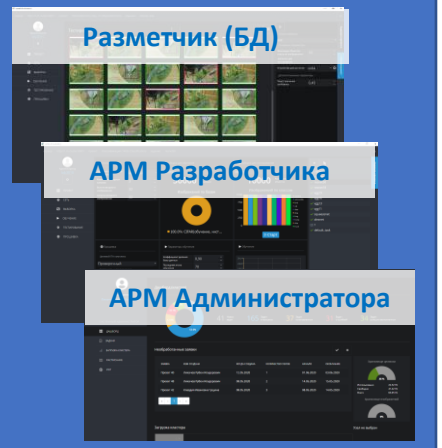
Программно-алгоритмические средства для решения задач анализа и синтеза многомодальных данных

Подр. Б.В.Вишнякова: Лаб. Ю.В.Морзеева, Лаб. М.А.Мельниченко

Унифицированная платформа нейросетевой разработки «Платформа-ГНС» (2018-2020)

Интегрированная среда

- Работа с данными
- Формирование и обучение ГНС
- Аппаратная реализация ГНС



Прикладные пакеты

- Готовые типовые решения
- Интегрированные решения
- Пользовательские решения

Типовые задачи

Компьютерное зрение

Анализ сигналов

Средства встраивания решений PlatformAPI

API

Унифицированная платформа

- **Сертифицируемый на НДВ исходный код**
- Импорт/экспорт, работа с фреймворками
- **Готовые решения для типовых задач**
- **Поддержка отечественного АО и ОС**
- Контроль доступа к данным и проектам
- **Низкие требования к квалификации ИТР**

Средства аппаратной реализации

Модуль

МЦСТ ЭЛЬБРУС

ЭЭЛВИС

Библиотека машинного обучения (фреймворк) PlatLib

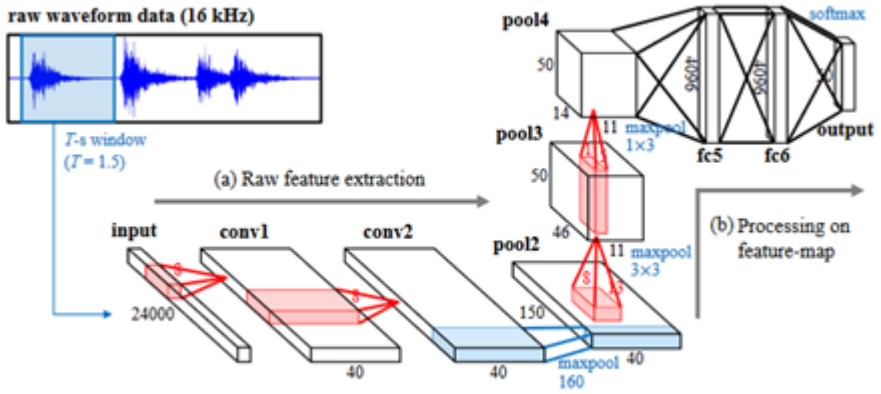
- Динамические графы
- Распределенное обучение
- Унификация с PyTorch

ГосНИИАС поддерживает и развивает полный отечественный программно-аппаратный стек технологий для ИИ

Проект «Автомат-ГНС» (2021-2023): синтез, аугментация, автообучение

СПО «Сигнал-ГНС» - набор типовых решений для обработки одномерных и многомерных сигналов с целью решения задач классификации, регрессии, устранения шумов.

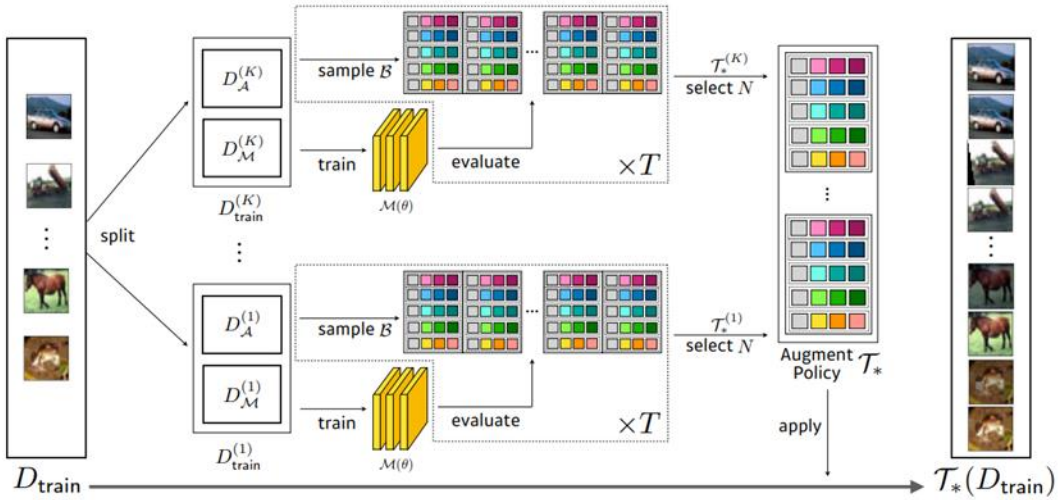
СПО «Авто-ГНС» — это набор программных модулей, которые позволяют пользователю в автоматическом режиме сформировать архитектуру ГНС (в задачах ТР) и подобрать оптимальные гиперпараметры обучения ГНС.



СПО «Аугмент-ГНС» – это набор программных модулей, которые используются в процессе обучения для определения оптимальной стратегии применения аугментаций (преобразований) к обучающей выборке.

СПО «Советник-ГНС» – это набор программных модулей, которые анализируют действия пользователя, формируют и выдают пользователю ситуативные рекомендации на этапах подготовки данных, обучения и тестирования.

СПО «Синтез-ГНС» – это набор программных модулей, которые реализуют возможность реалистичного нейростевого синтеза изображений и сигналов.



Планируемый проект «Фундамент-ГНС» (2024-26): MLLM + новые типы данных

Репозитории (БД)

БД данных для обучения и тестирования нейросетевых моделей (БД ДОТ)

БД аннотированных данных для обучения и тестирования нейросетевых моделей (БД АДОТ)

БД выборок данных для обучения и тестирования нейросетевых моделей (БД ВДОТ)

БД типовых решений (БД ТР);

БД пользовательских решений (БД ПР)

БД пользовательских проектов (БД ПП)

БД архитектур моделей ГНС (БД АМ)

БД обученных моделей ГНС (БД ОМ)

БД пайплайнов (схем) обработки данных при обучении, выполнении и тестировании ГНС (БД СОД)

БД модулей обучения и дообучения (БД МОД) для ПАК обучения и дообучения нейросетевых модулей конечных изделий (НМКИ)

СПО формирования, экспорта и импорта модулей обучения и дообучения (МОД) для ПАК обучения и дообучения бортовых НМКИ (СПО «Упаковщик-ГНС»)

СПО обучения и выполнения **больших генеративных языковых моделей** для работы с текстовыми данными (СПО «Модель-Я-ГНС»)

СПО дообучения и выполнения **фундаментальных моделей** для работы с многомодальными данными (СПО «Модель-Ф-ГНС»)

СПО **обучения нейросетевых моделей с подкреплением** (СПО «Подкрепление-ГНС»)

СПО обучения ГНС для **задач компьютерного зрения** (СПО «Зрение-ГНС») при работе с данными: ТВ и ИК изображениями; ТВ и ИК видеопоследовательностями; **РЛ; гидролокационными; лазерного 3D сканирования (облаками точек)**

СПО обучения ГНС для: **геоинформационных данных (СПО «Картограф-ГНС»); визуальной одометрии и автономной навигации (СПО «Навигатор-ГНС»); цифровых табличных данных (СПО «Таблица-ГНС»)**

СПО поддержки применения и информационного обеспечения САУ конечных изделий:

в интересах АО «КТРВ» (СПО «Изделие-КТР-ГНС»)

в интересах АО «КБМ» (СПО «Изделие-КБМ-ГНС»)

в интересах «ГНПП «Регион» (СПО «Изделие-РЕГ-ГНС»)

в части АО НПО «МИС» (СПО «Изделие-МИС-ГНС»)

Платформа-ГНС: Перспективы развития (дорожная карта)

Версия 1.7

Новые типы данных I

Автоматическое обучение I

Интеллектуальные помощники

Версия 2.0 «Автомат»

Абстрактное типовое решение

Облачное решение

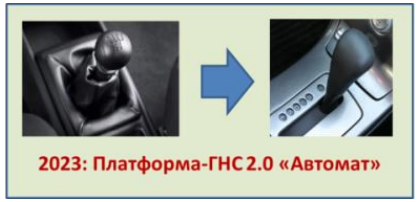
Новые типы данных II

Автоматическое обучение II

Новые прикладные пакеты RL, NLP,…

Большие и универсальные модели

Доступна для тестирования и использования с октября 2023



Документация, презентации, видеоматериалы

2022 **2023** **2024**

Многомодальные системы анализа и синтеза информации в задачах управления подвижными объектами

Визильтер Юрий Валентинович, д.ф.-м.н., проф. РАН, директор по направлению – руководитель научного комплекса «Искусственный интеллект и техническое зрение» ФАУ «ГосНИИАС», *viz@gosniias.ru*

Спасибо за внимание!

XIV Всероссийское совещание по проблемам управления (ВСПУ-2024)

Москва, ИПУ РАН, 20.06.2024