

年度重要学术进展-8

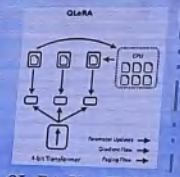
上海交通大学
沈为

大模型微调技术持续演进，大模型与小数据/小算力间的鸿沟不再难以跨越
VASE2024相关活动：APR-9(基础大模型)、Tutorial-3(开放词汇)、Workshop-5(智慧医疗)/9(大模型迁移)



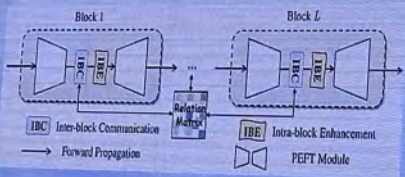
重要学术进展-8

小数据/小算力间的鸿沟不再难以跨越
(APR-9, Workshop-5(智慧医疗)/9(大模型迁移))



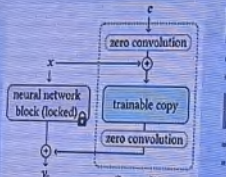
QLoRA (NeurIPS2023)

自然语言处理领域

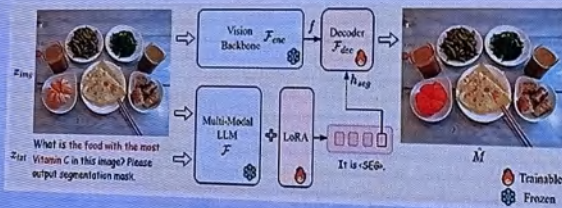


SAM-COBOT (CVPR2024)

计算机视觉领域



ControlNet (ICCV2023 best paper)



LISA (1.4K stars in 8 months)

多模态领域

年度重要学术进展-6

谢雨彤

德莱德大学

SAM出现并获得广泛应用, X Anything 范式开始流行

VALSE2024相关活动: APR-9(基础大模型)、Tutorial-3(开放词汇)、Workshop-9(大模型迁移)

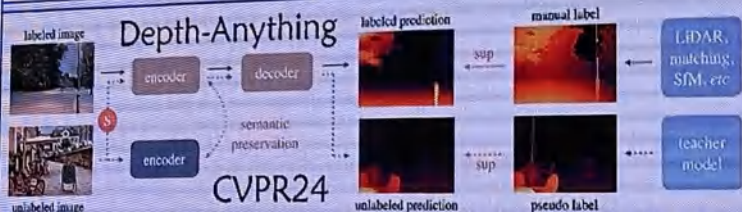
SAM被迅速应用至各种垂直域

- 医学影像分割
- 点云分割
- 遥感图像分割
- 全景图像分割
- 高效数据标注
- 视频超分辨率
- 视频目标追踪
- 机器人模型X
- 开放词汇目标检测
- 3D场景重建/理解
- 目标姿态估计
- 图像编辑
- 字幕生成
- 风格迁移
- 目标计数
-

- 改善SAM效率、稳定性和可扩展性
- BLO-SAM, ICML24
 - Conv-LoRA, ICLR24
 - MobileSAM
 - Finetune-SAM
 -

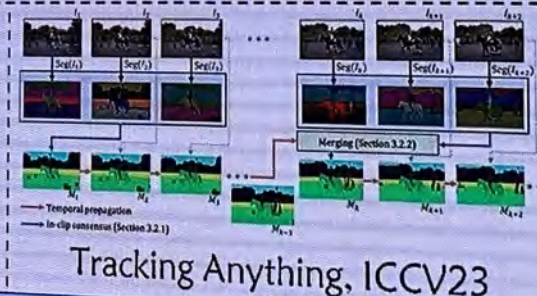
SAM

SAM (Segment Anything Model)
ICCV 2023 最佳论文提名



- Inpaint-Anything
- Matting-Anything
- Caption-Anything
- Anything-3D
- Ask-Anything
-

SAM成功推动了X Anything
研究范式的发展



年度重要学术进展-5

上海交通大学
陈思衡

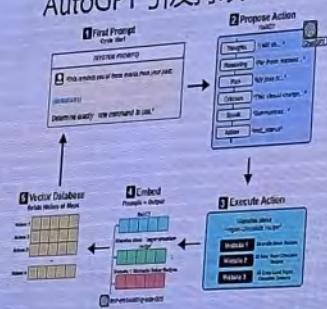
以AutoGPT为代表的AI agent范式开始流行，带来AI生产效率提升

VALSE2024相关活动：APR-2(智能体视觉)、Tutorial-1(具身智能与智能体)

AI小镇火遍全球，agent持续出圈

Aloha 2家务机器人效果拔群

AutoGPT 引发持续跟进



AutoGPT github开源
Google发布Palm-E具身
多模态模型

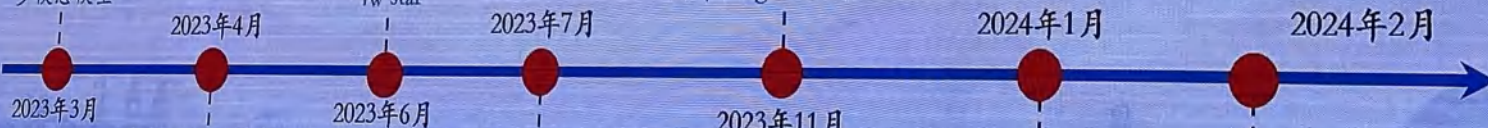


MetaGPT框架正式开
源，截至目前收获近
4w star

OpenAI上线GPT Shop低
代码平台GPT Builder，大
大降低agent搭建难度



微软上线Windows
agent框架UFO，两个
月时间收获4K stars



斯坦福AI小镇发表
“Generative agents: Interactive
simulacra of human behavior”

CMU WebArena 框架发布
WebAgent迎来大发展

李飞飞团队发布综述：
“Agent AI: Surveying the Horizons of
Multimodal Interaction”

斯坦福团队发布Aloha 2家
务机器人，引发全网热议

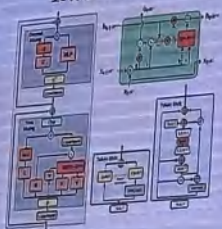
从纯语言到多模态到具身智能，从“做一件事”到“让AI做一件事”：学术界/产业界/创投界正以空前的热情投入AI agent的研究与发展

年度重要学术进展-4

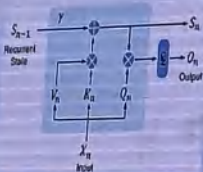
华中科技大学
王兴刚

高效大语言模型持续发展, 新型CV计算架构持续涌现
VALSE2024相关活动: APR-8(新型高效网络架构)/9(基础大模型)、Workshop-12(大模型理论与机理)

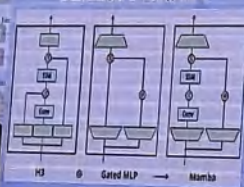
RWKV Network



Retention Network

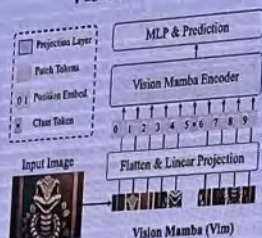


Mamba Network

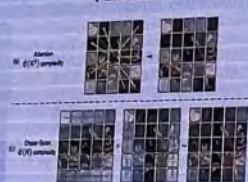


高效大语言模型

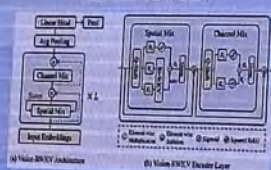
Vision Mamba



VMamba

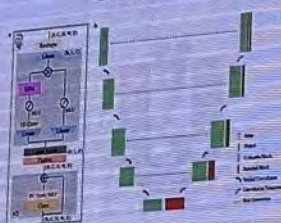


Vision RWKV



骨干网络

U-Mamba



VM-UNet

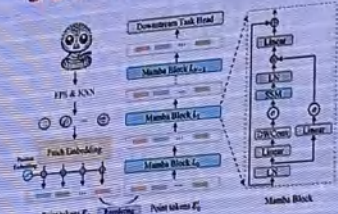


MambaIR / VmambaIR

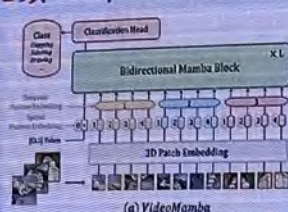


医学图像&底层视觉

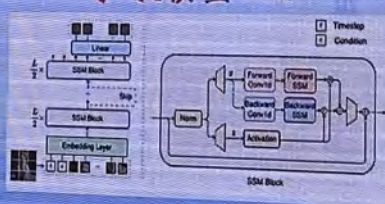
3D点云 PointMamba



视频理解 VideoMamba

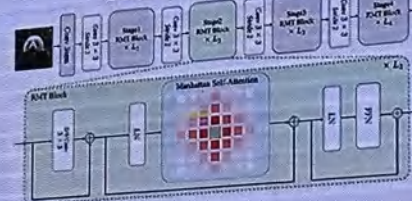


扩散模型 DiS

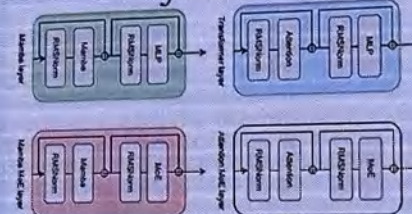


视觉理解&图像生成

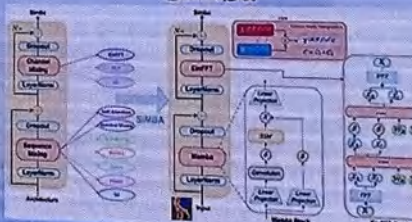
RMT Network



Jamba



Simba



混合架构&多模态

年度重要学术进展-3

北京大学
王鹤

具身视觉在大模型加持下显著扩展其范畴, CV算法与环境的交互愈发丰富
VAISE2024相关活动: APR-5(具身感知交互)/11(具身决策), Tutorial-1(具身智能), Workshop-4(具身智能)

大脑



PaLM-E - Google



GPT-4V - OpenAI



LLM-Planer [ICCV23]



EmbodiedGPT [NeurIPS23]

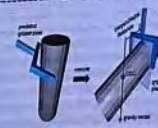


Voxposer [CoRL23]



SAGE

小脑



AnyGrasp [TRO23]



RT-2 [CoRL23]



GR-1 [ICLR23]



RoboCook [CoRL23 best system paper]



Diffusion Policy [RSS 23]



Eurek



UniDexGrasp++ [ICCV23 best paper finalist]

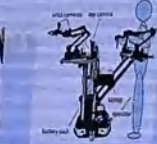


NVM [CVPR 23 Highlight]

本体



LEAP Hand



Mobile Aloha



Hello Robot Galbot



1X



Tesla



Digit



Fouri



Unitree H1



New Atlas



Figure 01

数据集



RoboGen



Open X-Embodiment



ManiSkill2



DexGraspNet [ICRA 23 outstanding manipulation finalist]



EmbodiedScan [CVPR24]



DROID

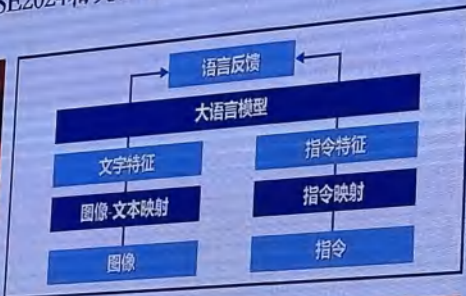
鵬城實驗室

楊文瀚

年度重要学术进展-2

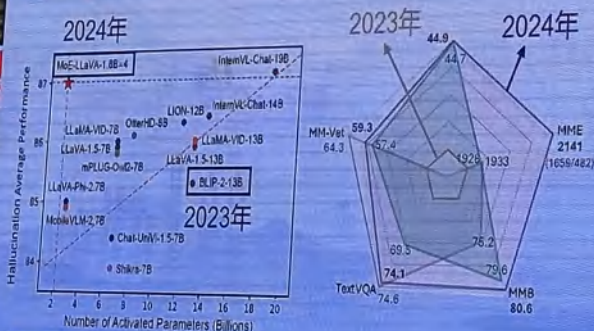
视觉与学习青年学者研讨会
Valse VISION AND LEARNING SEMINAR

多模态大语言模型的进化，带来多模态对话新范式，扩展了视觉理解的外延
Valse2024相关活动：APR-5(多模态感知交互)、Workshop-13(多模态感知对话)/18(多模态大模型)



以大语言模型为抓手，实现跨模态对话问答

多模态对话系统百花齐放



性能持续进步

图文对话

1. There is a solution to a math problem by a student. Try to reason about the question step by step. Did the student get the correct answer?

The student has attempted to solve the math problem by adding the number of minutes. Each practice on Monday and Tuesday.

The student has correctly added 254 minutes (254 is 208 + 46) to 133 minutes, resulting in 387 minutes. However, the student has not accounted for the fact that each practice session goes on Monday and Tuesday.

The correct answer should be the difference between the total minutes (387 minutes) practiced on Monday and Tuesday, which is 384 minutes (Monday: 133 minutes (Tuesday) = 121 minutes. Therefore, after practicing 133 more minutes on Monday than on Tuesday, the student's answer is 133 more minutes than the Tuesday practice time from the total Monday practice time.

OCR

Feature	13-inch MacBook Pro	15-inch MacBook Pro
Processor	Intel Core i5	Intel Core i7
Graphics	Intel Iris Graphics 6100	Intel Iris Pro Graphics
Battery Life	Up to 10 hours	Up to 8 hours
Trackpad	Force Touch trackpad	Multi-Touch trackpad
Speed	2.7GHz, 2.9GHz	2.2GHz, 2.5GHz
Memory	8GB	16GB
Storage	128GB, 256GB, 512GB	256GB, 512GB
Price	\$1,699.00, \$1,999.00, \$2,399.00	\$2,699.00, \$3,299.00

图像生成

Capture a scene of space elevator

Media saying AI will take over the world

My Neural Network

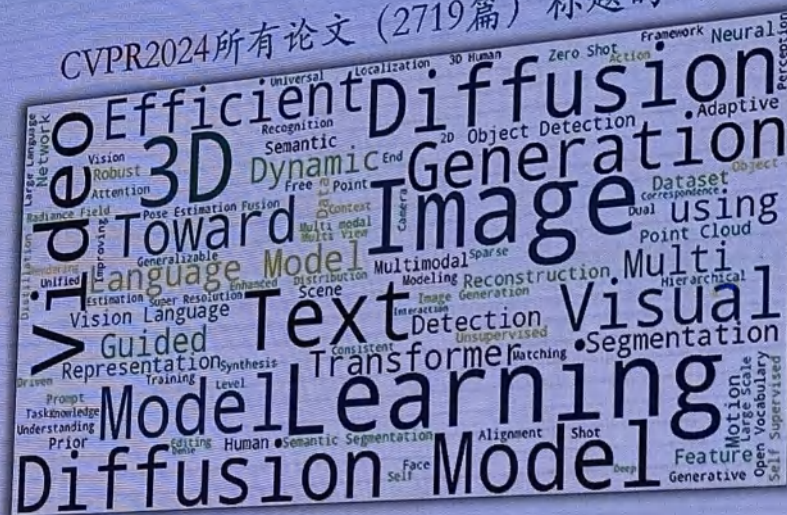
图像理解

5. Explain this meme and illustrate what it implicates.

The meme is a humorous take on the common narrative that artificial intelligence (AI) will become so advanced that it will eventually surpass human capabilities and perhaps even take over the world. The left panel shows a person reacting with alarm, which is a common portrayal of the public's reaction to neural networks, which is a humorous example of AI making a mistake. This juxtaposition implies that while AI is indeed becoming more advanced, it is not yet at the level where it can accurately perform tasks such as image recognition without errors. The meme uses this contrast to humorously downplay the fear of AI taking over, suggesting that the technology is still in its early stages and needs more guidance.

涌现出在图文对话、图像理解、OCR等方面的卓越能力

CVPR2024所有论文 (2719篇) 标题的词云



消逝的过往

兴起的未来

- 传统任务: Object Detection
- 传统数据: Point Cloud
- 模型崇拜: Transformer

- 结合语言: Language Model
- 生成任务: Generation
- 生成模型: Diffusion Model

年度重要学术进展-1

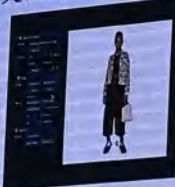
北京大学
袁粒

学术
进展

高清图像和视频生成技术快速发展，基础模型涌现物理世界建模能力
VALSE2024相关活动：APR-4(视频生成)、Tutorial-2(视频生成)、Workshop-3(视频生成)

OpenAI提出
Consistency Models
一步生成扩散模型

ControlNet(斯坦福)和
T2I-Adapter(北大)分别被
提出，用于文生图精准控制



德国马普所、MIT
等提出点追踪算法
的DragGAN模型

港中文和上海AI
Lab共同提出
AnimateDiff文生
动态GIF模型



华为基于Diffusion
Transformer提出文
生图模型PixArt- α

阿里提出图生视频
Animate Anyone生成人
物角色动画模型

Stability AI公开论
文SD3，基于多模态
DiT和Flow Model
的文生图模型

阿里达摩院提出
EMO，图像和音乐
生成人物画像视频
模型

莫纳什大学和上海AI
Lab等基于DiT提出
2D+1D attention文
生视频模型Latte

2023.02-03

2023.04-06

2023.07-10

2023.10-12

2024.01-04

Runway发布第一
版商业文生视频应
用Gen-1

Midjourney V5系
列发布，文生图质
量和美学新高度

OpenAI推出文
生图产品DALL-
E3，与ChatGPT
深度融合

Pika发布文生视频工
具Pika 1.0，
Runway发布Gen-2



Stability AI正式公布
最新的开源绘图应用—
SDXL1.0

Stability AI开
源Stable Video
Diffusion

OpenAI放出Sora文生
视频模型引发巨大关注，
完全闭源仍无API接口
和使用通道

生数科技推出Vidu但未
开源，北大和NUS团队
分别发起Sora复现计划

应用
进展