Introduction of Ego4D and Ego4D Challenges

Guo Chen

Nanjing University & Shanghai AI Lab

Outline

- Background of Ego4D Dataset
- Benchmark Tasks for Ego4D
- Ego4D Challenge
- Our solution on ECCV2022

Outline

- **Background of Ego4D Dataset**
- Benchmark Tasks for Ego4D
- Ego4D Challenge
- Our solution on ECCV2022

Motivation: Previous Datasets

- 1）从第三视角以旁观者的角度捕捉短暂、孤立的时刻
- 2）之前的大部分图像或者视频的数据集，都是有意拍摄的
- 3）难以从人类的视角去理解更高级别的行为

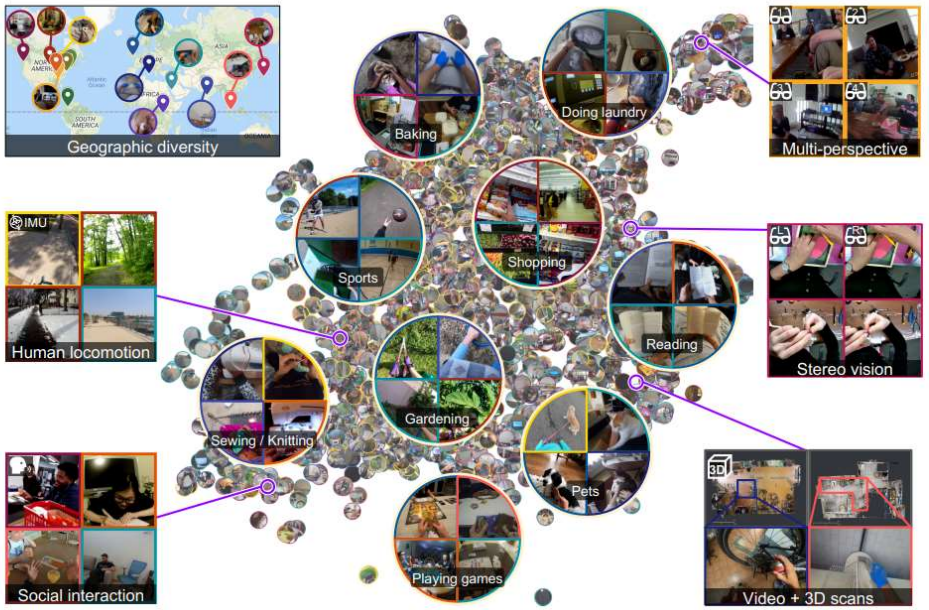
Motivation: Ego4D Datasets

- 1）从第三视角以旁观者的角度捕捉短暂、孤立的时刻
  机器人和AR技术中，更应该关注第一视角的长时且流畅的视频流

- 2）之前的大部分图像或者视频的数据集，都是有意拍摄的
  第一视角相机不会有意拍摄特定的视频片段，而是一个存在自我意识的视频流

- 3）难以从人类的视角去理解更高级别的行为
- 第一视角相机需要保持一个持久的3D场景理解能力，从人类的角度去理解每一个高级行为，例如人和物体的交互，人和人的社交行为。

Ego4D Datasets



931名参与者来自9个国家/地区的74个地点拍摄了共3670小时的第一视角视频。

数据非特意拍摄：大部分镜头无脚本是在野外自然拍摄，代表拍摄者在工作生活中的自然交互。

自我认同机制：拍摄者来源于不同背景、职业、性别和年龄的人。

地理多样：丰富的地理环境保证了交互对象，交互行为的多样性，也补充了之前大多数数据集不存在的对象和行为。

时序持久：拍摄1到10小时的视频保证长时视频流。

模态多样：除了RGB视频，部分视频还引入了音频、3D mesh、Gaze等模态。

Outline

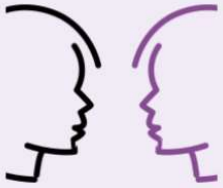
- Background of Ego4D Dataset
- **Benchmark Tasks for Ego4D**
- Ego4D Challenge
- Our solution on ECCV2022

Five Benchmark tasks

Episodic Memory

Hand-Object Interactions

AV Diarization



Social

Forecasting

Five Benchmark tasks

- 情节记忆：包含了一系列1D和2D的检测任务
- 手部目标交互：包含了一系列对于正在交互的物体的状态改变的检测
- 视听诊断和社交：包含了一些视频和音频结合的任务，包括1D和2D的定位和语音识别等。
- 未来预测：包含了一些预测未来信息的任务。

Outline

- Background of Ego4D Dataset
- Benchmark Tasks for Ego4D
- **Ego4D Challenge**
- Our solution on ECCV2022

CVPR2022

Episodic Memory

- Visual queries 2D localization (VQ)
- Visual queries 3D localization (VQ3D)
- Natural language queries (NLQ)
- Moments queries (MQ)

Hands and Objects

- PNR localization (PNR)
- State change classification (OSCC)
- State change object detection (SCOD)

Audio-Visual & Social

- AV localization (AVLoc)
- Audio-only diarization (ADiar)
- Audio-visual diarization (AVDiar)
- AV Speech transcription (AVTrans)
- Talking to me (TTM)
- Looking at me (LAM)

Forecasting / Anticipation

- Future hand prediction (FHP)
- Short-term anticipation (STA)
- Long-term anticipation (LTA)

ECCV2022

Episodic Memory

- Visual queries 2D localization (VQ)
- Visual queries 3D localization (VQ3D)
- Natural language queries (NLQ)
- Moments queries (MQ)

Hands and Objects

- PNR localization (PNR)
- State change classification (OSCC)
- State change object detection (SCOD)

Audio-Visual & Social

- AV localization (AVLoc)
- Audio-only diarization (ADiar)
- Audio-visual diarization (AVDiar)
- AV Speech transcription (AVTrans)
- Talking to me (TTM)
- Looking at me (LAM)

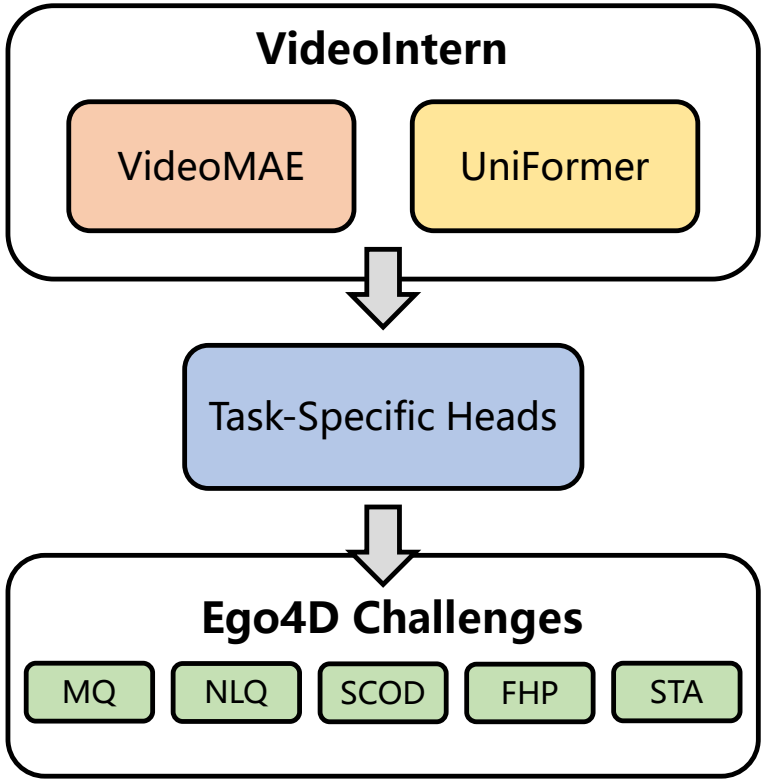
Forecasting / Anticipation

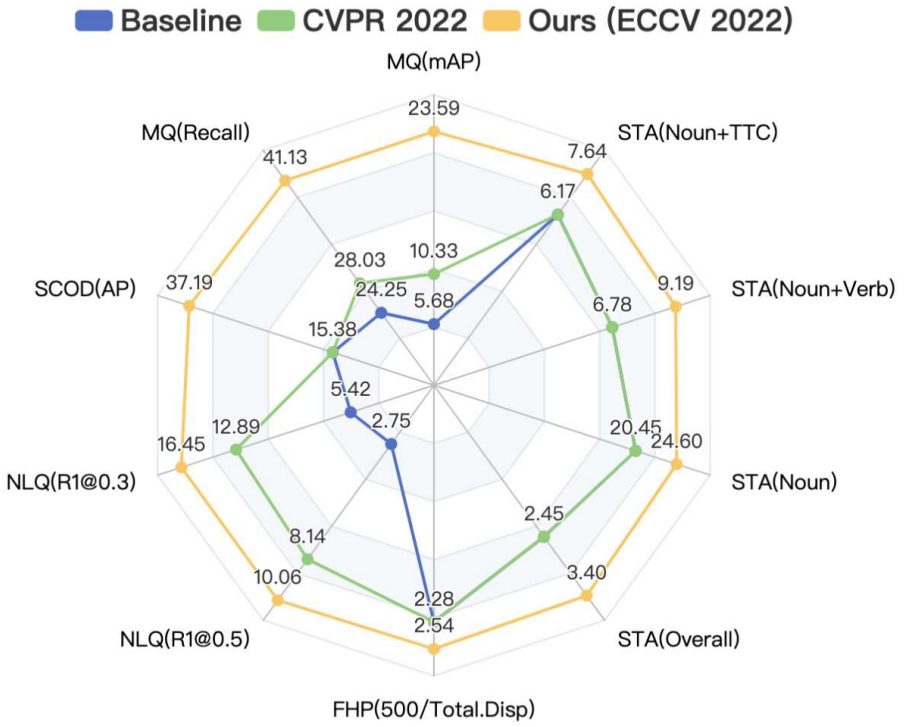
- Future hand prediction (FHP)
- Short-term anticipation (STA)
- Long-term anticipation (LTA)

Outline

- Background of Ego4D Dataset
- Benchmark Tasks for Ego4D
- Ego4D Challenge
- **Our solution on ECCV2022**

Baseline: InternVideo





Thanks

Technical Report: *InternVideo-Ego4D: A Pack of Champion Solutions to Ego4D Challenges*
Code: *https://github.com/OpenGVLab/ego4d-eccv2022-solutions*