NEW YORK UNIVERSITY Meta AI

目标驱动的人工智能

迈向能够学习、记忆、推理、计划、有常识，但可操纵且安全

杨立昆Yann LeCun
纽约大学
Meta – 基础 AI 研究

Ding-Shum讲座 哈佛大学数学系
2024-03-28


M-51, HSO

机器学习很糟糕！（与人类和动物相比）

- **监督学习 （SL） 需要大量标记样本。**
- **强化学习 （RL） 需要大量的试验。**
- **自我监督学习 （SSL） 效果很好，但是……**
  - 生成式预测仅适用于文本和其他离散模态

动物和人类：

- 可以非常快速地学习新任务。
- 了解世界是如何运作的
- 可以推理计划

人类和动物都有常识
那里的行为是由目标（驱动器）驱动的

我们需要人类级别的人工智能来智能助手

- **在不久的将来，我们与数字世界的所有互动都将由人工智能助手进行调解。**
- **智能眼镜**
  - 通过语音、视觉、显示、肌电图接口 （EMG） 进行通信
- **智能辅助**
  - 可以回答我们所有的问题
  - 可以在日常生活中帮助我们
  - 了解我们的喜好和兴趣
- **为此，我们需要具有人类水平智能的机器**

“Her”
(2013)

- 了解世界如何运作的机器
- 能够记忆、推理、计划的机器。

未来的 AI 助手需要人类水平AI

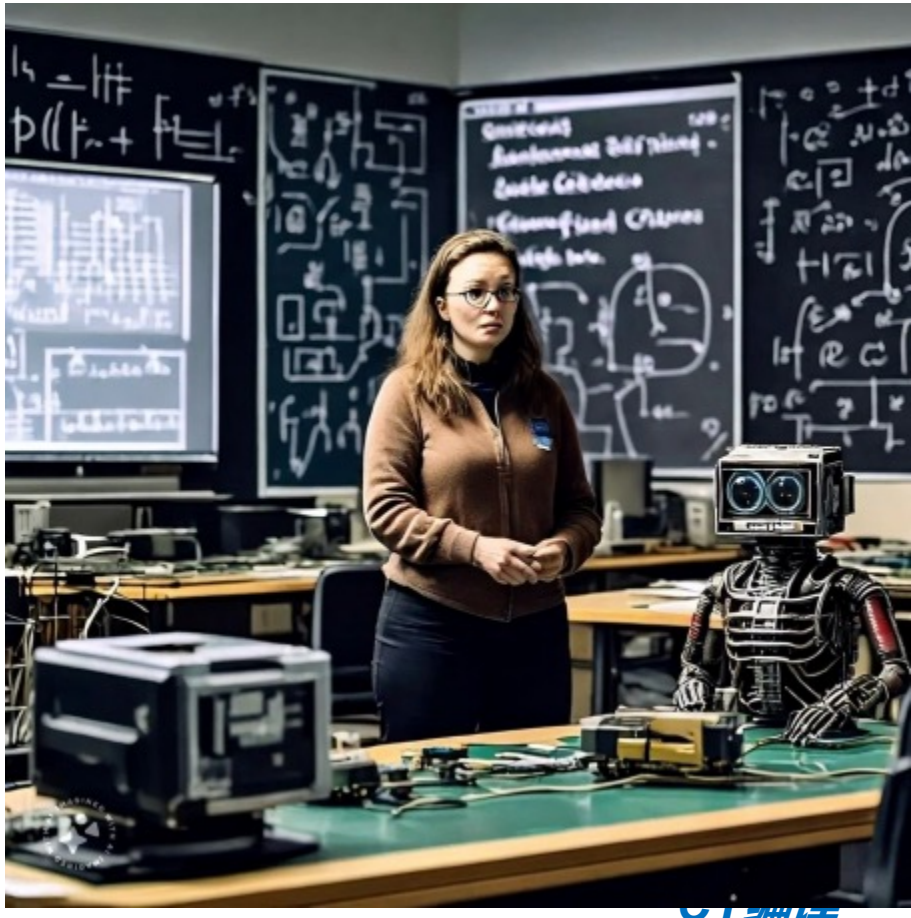
- 人工智能助手将需要（超级）人类水平的智能
  - 就像有一群聪明的“人”为我们工作一样

- 但是，我们今天远未达到人类水平的人工智能
  - 任何 17 岁的孩子都可以在 20 小时的训练中学会驾驶
  - 任何 10 岁的孩子都可以学会一口气清理餐桌
  - 任何家猫都可以计划复杂的行动

- 我们错过了什么?
  - 学习如何世界工作（不仅仅是从文本中）世界模型。
  - 常识
  - 记忆、推理、分层规划

面向 AMI（高级机器智能）的 Desiderata

- **从感官输入中学习世界模型的系统**
  - E.g.从视频中学习直观的物理原理
- **具有持久记忆的系统**
  - 大规模联想记忆
- **可以计划行动的系统**
  - 从而实现一个目标
- **可控和安全的系统**
  - 通过设计，而不是通过微调。
- **目标驱动的 AI 架构**

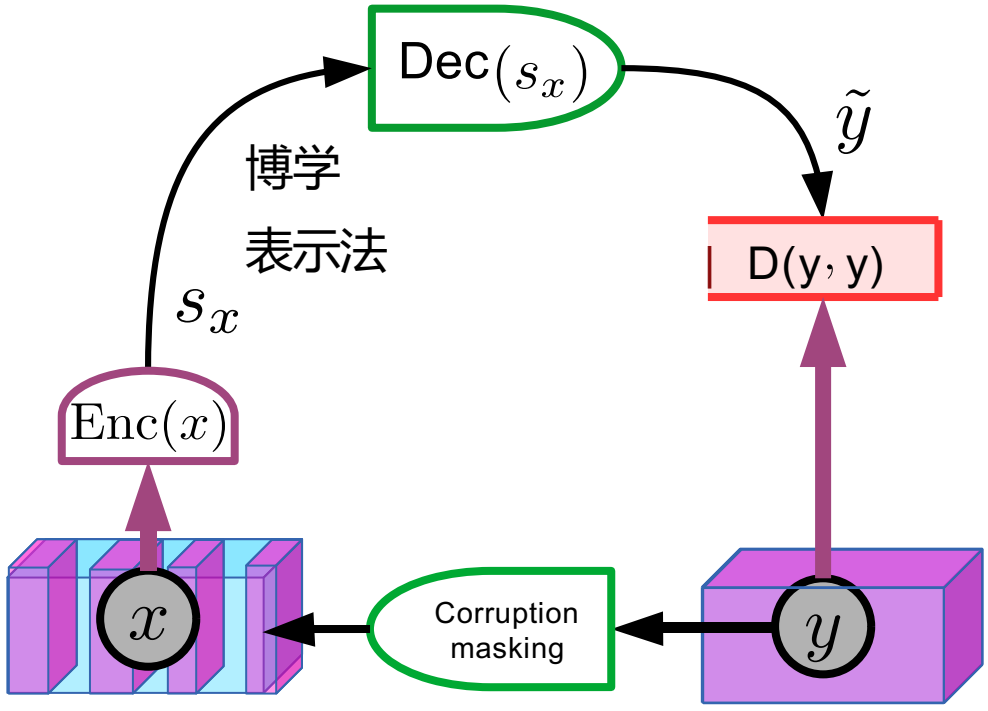


自我监督学习
已经占领了世界

用于理解和生成文本、图像、视频、3D模型、语音、蛋白质。。。

通过去噪/重构进行自我监督学习

- 去噪自动编码器 [Vincent 2008]、BERT [Devlin 2018]、RoBERTa [Ott 2019]



这是摘录的文本[...]，是一大批[...]文章

这是从大量新闻文章中摘录的一段文字

Emu:图像生成

[ArXiv:2309.15807]

Dai 等人：

鸸鹋：增强图像生成

大海捞针的模型

2023 年 9 月，Meta 的 AI

WhatsApp 和 Messenger 上的 Meta AI：

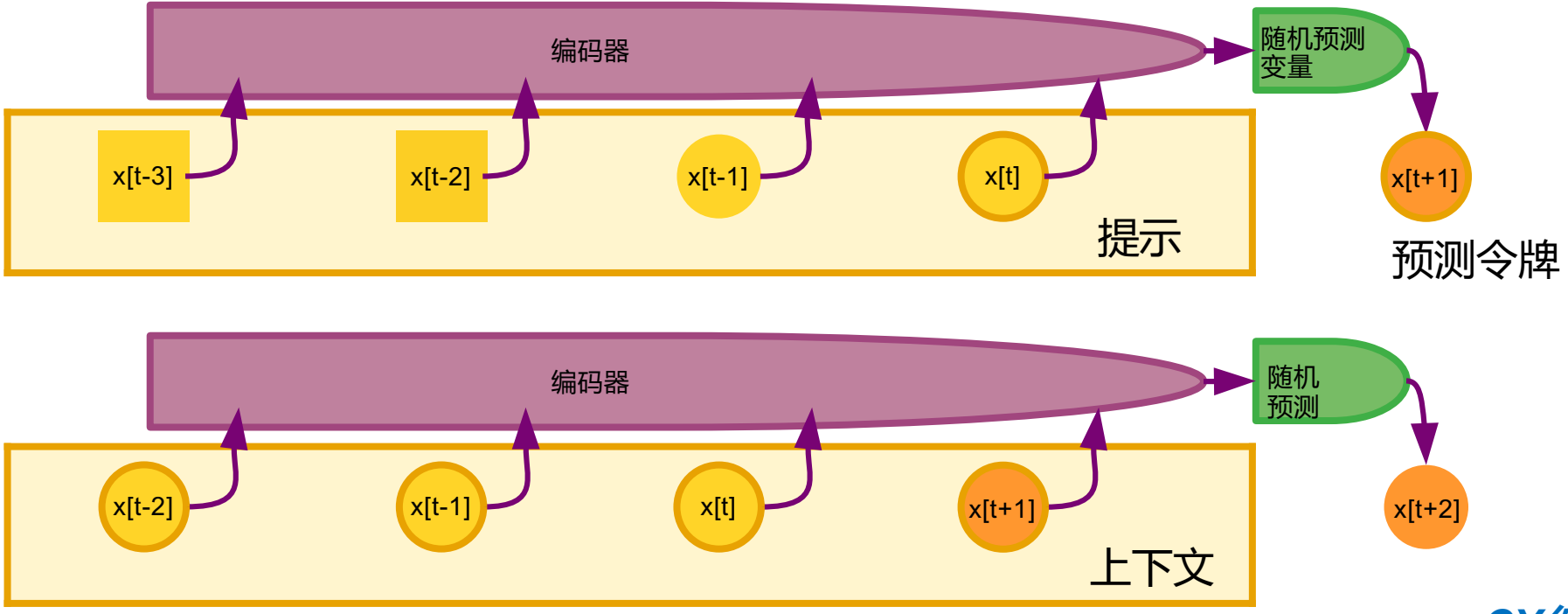
/想象一张照片是哈佛

数学家在智能机器人的帮助下在黑板上证明了黎曼假设。

生成式 AI 和
自动回归
大型语言模型

自回归生成式架构

- 输出一个又一个的“令牌”
- 标记可以代表单词、图像补丁、语音片段……



自回归大型语言模型 （AR-LLMs））

- **一个接一个地输出文本标记**
- **标记可以表示单词或子单词**
- **编码器/预测器是一种变压器架构**
  - 具有数十亿个参数：通常从 1B 到 500B
  - 训练数据：1 到 2 万亿个Tokens
- **用于生成对话框/文本的 LLM：**
  - 开放： BlenderBot， Galactica， LlaMA， Llama-2， Code Llama （FAIR）， Mistral-7B， Mixtral-4x7B （Mistral）， Falcon （UAE）， Alpaca （Stanford）， **Yi** **(**01.AI）， OLMo （AI2）， Gemma （Google） ....
  - 专有：Meta AI （Meta）、LaMDA/Bard、Gemini （Google）、ChatGPT （OpenAI）…
- **性能令人惊叹.....但。。。他们犯了愚蠢的错误**
  - 事实错误、逻辑错误、不一致、推理有限、毒性.....
- **LLM对潜在现实的了解有限**
  - 他们没有常识，没有记忆，他们无法计划他们的答案

CY编译

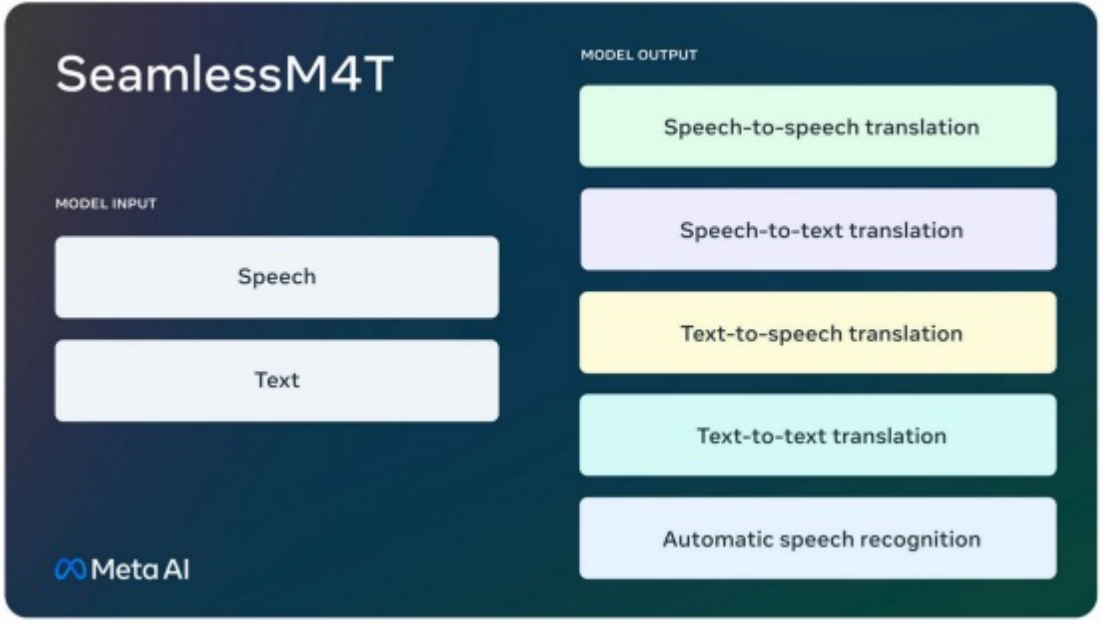
Llama-2: https://ai.meta.com/llama/

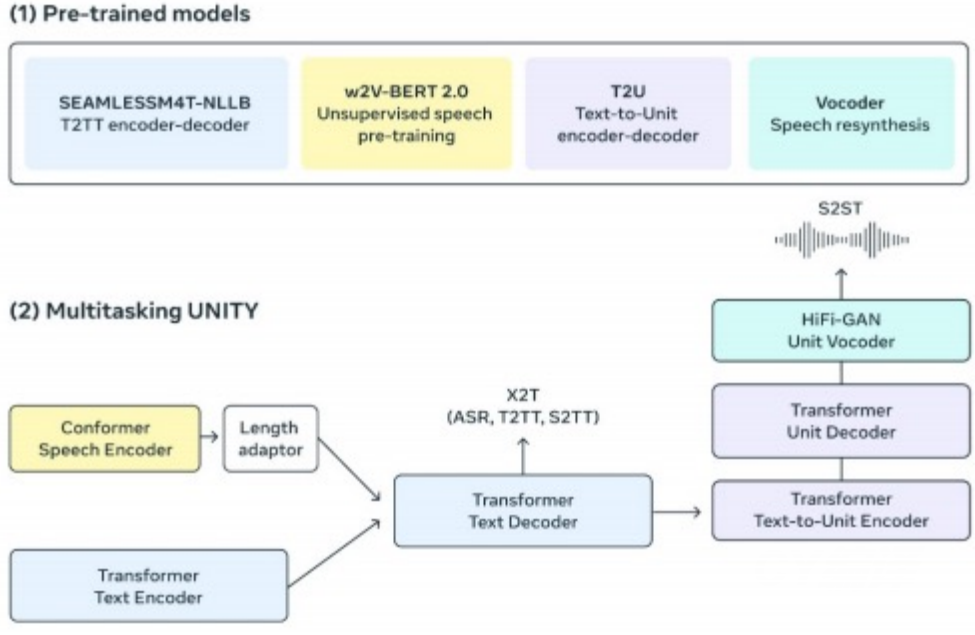
- 开源代码 / 免费 & 开放模型 / 可以商业使用
- 在 Azure、AWS、HuggingFace 上可用, ...

MODEL SIZE (PARAMETERS)	PRETRAINED	FINE-TUNED FOR CHAT USE CASES
7B	Model architecture:	Data collection for helpfulness and safety:
13B	Pretraining Tokens: 2 Trillion	Supervised fine-tuning: Over 100,000
70B	Context Length: 4096	Human Preferences: Over 1,000,000

无缝M4T

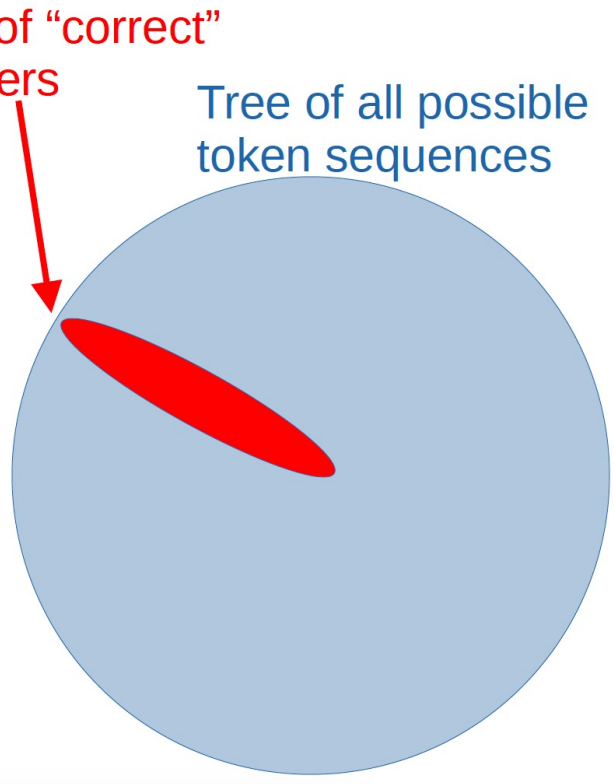
- 语音或文本输入：100 种语言
- 文本输出：100种语言
- 语音输出：35 种语言
- 无缝表达：实时，保留语音和表情
- https://ai.meta.com/blog/seamless-m4t/





自回归生成模型很糟糕!

- Auto-Regressive LLMs are **doomed**.
- They cannot be made factual, non-toxic, etc.
- They are not controllable
- Probability e that any produced token takes us outside of the set of correct answers
- Probability that answer of length n is correct:
  - $P(\text{correct}) = (1-e)^n$
- **This diverges exponentially.**
- **It's not fixable (without a major redesign).**

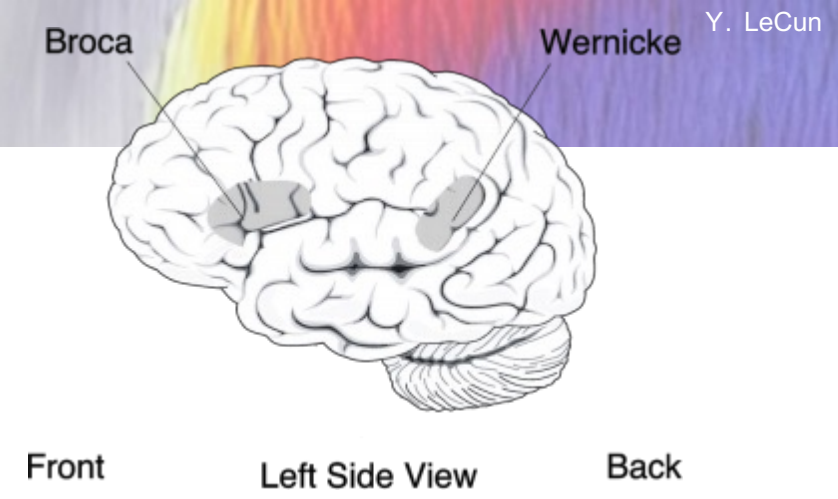


- See also [Dziri...Choi, ArXiv:2305.18654]

LLM 的局限性：没有规划!

- 自动回归 LLM（充其量）近似函数
  大脑中的 Wernicke 和 Broca 区域.

- 前额叶皮层呢?



ArXiv:2301.06627

ArXiv:2206.10498

DISSOCIATING LANGUAGE AND THOUGHT IN LARGE LANGUAGE MODELS: A COGNITIVE PERSPECTIVE

A PREPRINT

Kyle Mahowald*
The University of Texas at Austin
mahowald@utexas.edu

Anna A. Ivanova*
Massachusetts Institute of Technology
annaiv@mit.edu

Idan A. Blank
University of California Los Angeles
iblank@psych.ucla.edu

Nancy Kanwisher
Massachusetts Institute of Technology
ngk@mit.edu

Joshua B. Tenenbaum
Massachusetts Institute of Technology
jbt@mit.edu

Evelina Fedorenko
Massachusetts Institute of Technology
evelina9@mit.edu

Large Language Models Still Can't Plan (A Benchmark for LLMs on Planning and Reasoning about Change)

Karthik Valmeekam*
School of Computing & AI
Arizona State University, Tempe.
kvalmeek@asu.edu

Alberto Olmo*
School of Computing & AI
Arizona State University, Tempe.
aolmo@asu.edu

Sarath Sreedharan †
Department of Computer Science,
Colorado State University, Fort Collins.
sarath.sreedharan@colostate.edu

Subbarao Kambhampati
School of Computing & AI
Arizona State University, Tempe.
rao@asu.edu

自回归生成模型很糟糕！

- **AR-LLMs**
  - 在输入和输出之间具有恒定数量的计算步骤。代表性较弱。
  - 不要真的讲道理。没有真正的计划，没有常识
- **Noema 杂志，2023 年 8 月**

AI And The Limits Of Language

An artificial intelligence system trained on words and sentences alone will never approximate human understanding.

ESSAY TECHNOLOGY & THE HUMAN

BY JACOB BROWNING AND YANN LECUN

AUGUST 23, 2022

自动回归 LLM 糟透了！

- **自动回归 LLM 适用于**
  - 写作协助，初稿生成，文体润色。
  - 代码编写帮助
- **他们不好的地方：**
  - 提供事实和一致的答案(hallucinations!)
  - 考虑到最近的信息（在上次培训之前）
  - 行为正确（它们模仿训练集中的行为）
  - 推理、计划、数学
  - 使用“工具”，例如搜索引擎、计算器、数据库查询.....
- **我们很容易被他们的流利程度所愚弄。**
- **但他们不知道世界是如何运作的。**

目前的人工智能技术（仍然）与人类水平相去甚远

- **机器不会像动物和人类那样学习世界的运作方式**
- **自动回归 LLM 无法接近人类水平的智能**
  - 流利，但有限的世界模型，有限的计划，有限的推理。
  - 大多数人类和动物的知识都是非语言的。
- **我们仍然缺少在实现动物智能方面的重大进展**
  - 人工智能在某些狭窄的领域是超人

- **毫无疑问，最终，机器将在所有领域超越人类智能**

  - 人类的总智慧将会增加
  - 我们应该对此表示欢迎，而不是害怕它。

我们错过了一些真正重要的东西！

- **没关系，人类，猫和狗可以做出惊人的壮举**
  - 机器人智能远不及什么
- **任何 10 岁的孩子都可以学会在几分钟内清理餐桌并装满洗碗机。**
  - 我们没有可以做到这一点的机器人。
- **任何 17 岁的孩子都可以在 20 小时的练习中学会驾驶汽车**
  - 我们仍然没有无限的 Level-5 自动驾驶
- **任何家猫计划复杂的行动**
- **我们不断遇到莫拉维克的悖论**
  - 对人类来说容易的事情对人工智能来说很难，反之亦然。

数据带宽和容量：LLM 与孩子。

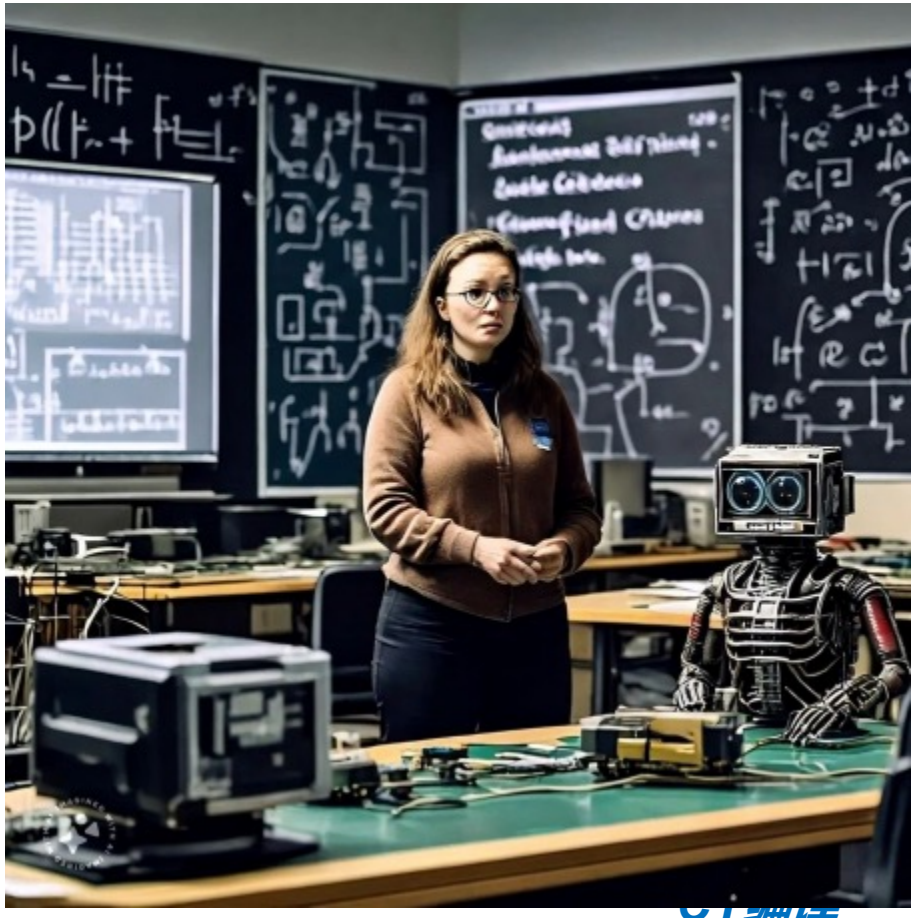
- **LLM**
  - 使用 1.0E13 令牌（0.75E13 个单词）进行训练。每个令牌为 2 个字节。
  - 数据量：2.0E13 字节。
  - 人类需要 170,000 年才能阅读（8 小时/天，250 w/分钟）

- **人类孩子**
  - 前 4 年唤醒 16,000 小时（YouTube 上传 30 分钟）
  - 200 万根视神经纤维，每根携带约 10 字节/秒。
  - 数据量：1.1E15 字节

- **一个四岁的孩子看到的数据是法LLM的 50 倍！**
- **在 300 小时内，孩子看到的数据比 LLM 还多。**

我们错过了什么？

- **从感官输入中学习世界模型的系统**
  - 例如，从视频中学习直观的物理知识
- **具有持久记忆的系统**
  - 大规模联想记忆
- **可以计划行动的系统**
  - 从而实现一个目标
  - 像人类“系统2”一样的理性
- **可控和安全的系统**
  - 通过设计，而不是通过微调.
- **目标驱动的 AI 架构**



目标驱动的人工智能系统

可以学习、推理、计划、

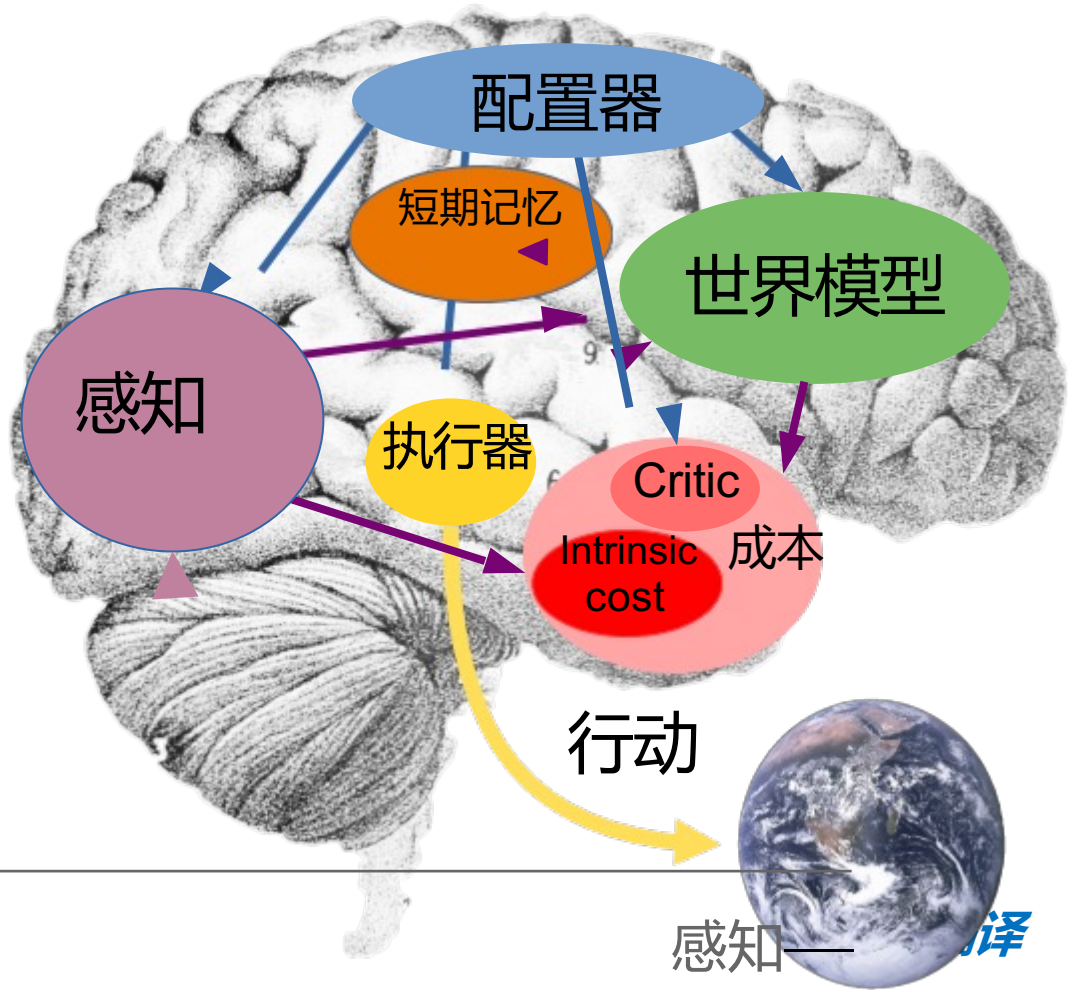
但安全可控

“通往自主机器智能的道路”
https://openreview.net/forum?id=BZ5a1r-kVsf

[YouTube上此演讲的各种版本]

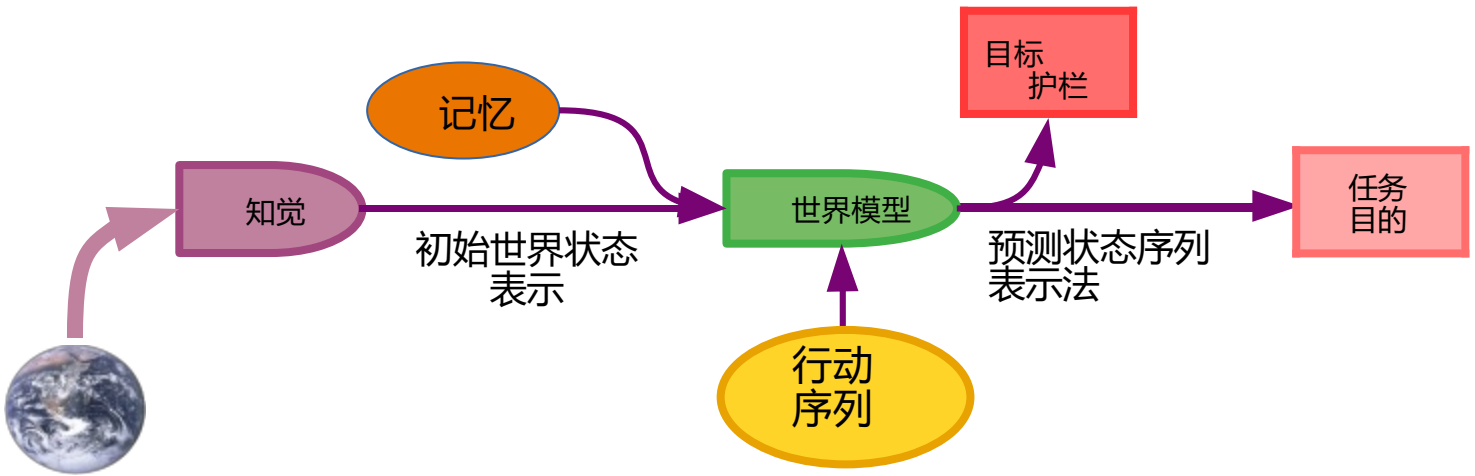
用于目标驱动型 AI 的模块化认知架构

- **配置器**
  - Configures other modules for task
- **感知**
  - Estimates state of the world
- **世界模型**
  - Predicts future world states
- **成本**
  - Compute “discomfort”
- **执行器**
  - Find optimal action sequences
- **短期记忆**
  - Stores state-cost episodes



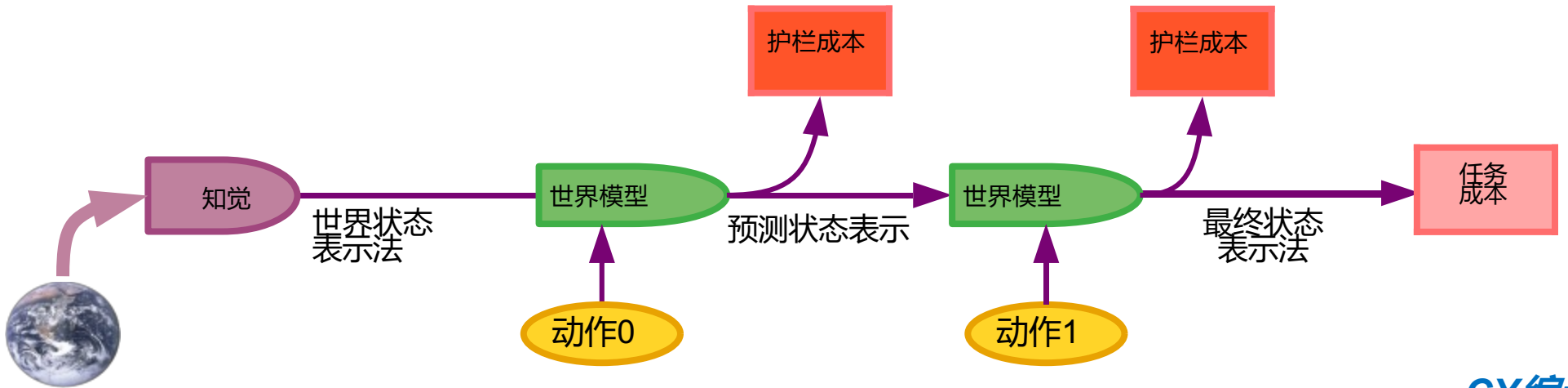
目标驱动的人工智能

- **感知：计算世界状态的抽象表示**
  - 可能与内存中先前获取的信息相结合
- **世界模型：预测由想象的动作序列产生的状态**
- **任务目标：衡量与目标的背离**
- **护栏目标：确保安全的不可变客观术语**
- **操作：查找最小化目标的操作序列**



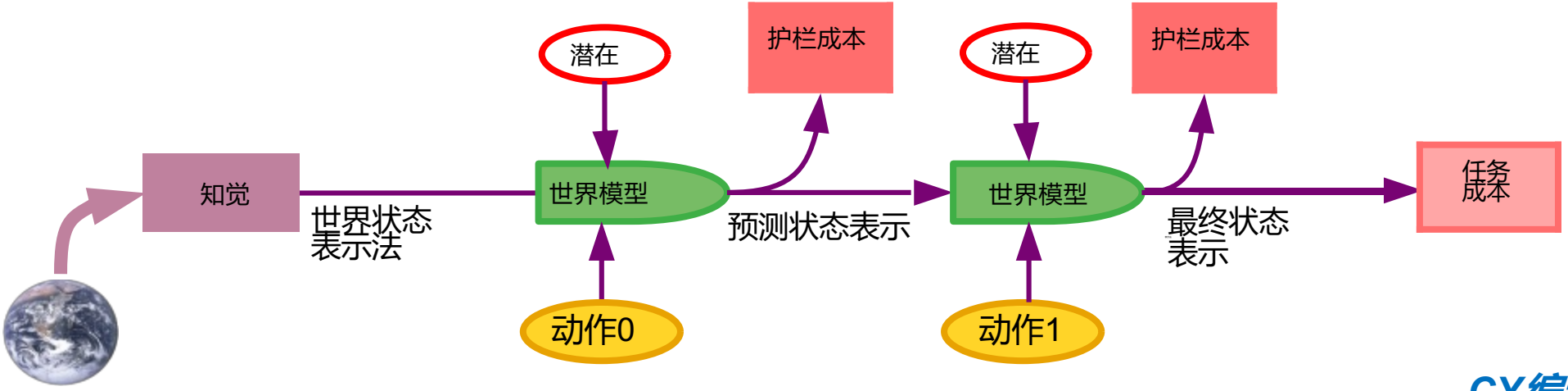
目标驱动的人工智能：多步骤/循环世界模型

- 在多个时间步长应用的同一世界模型
- 应用于整个状态轨迹的护栏成本
- 这与模型预测控制 （MPC） 相同
- 通过最小化目标进行行动推断
  - 使用基于梯度的方法、图形搜索、动态 prog、A*、MCTS、....



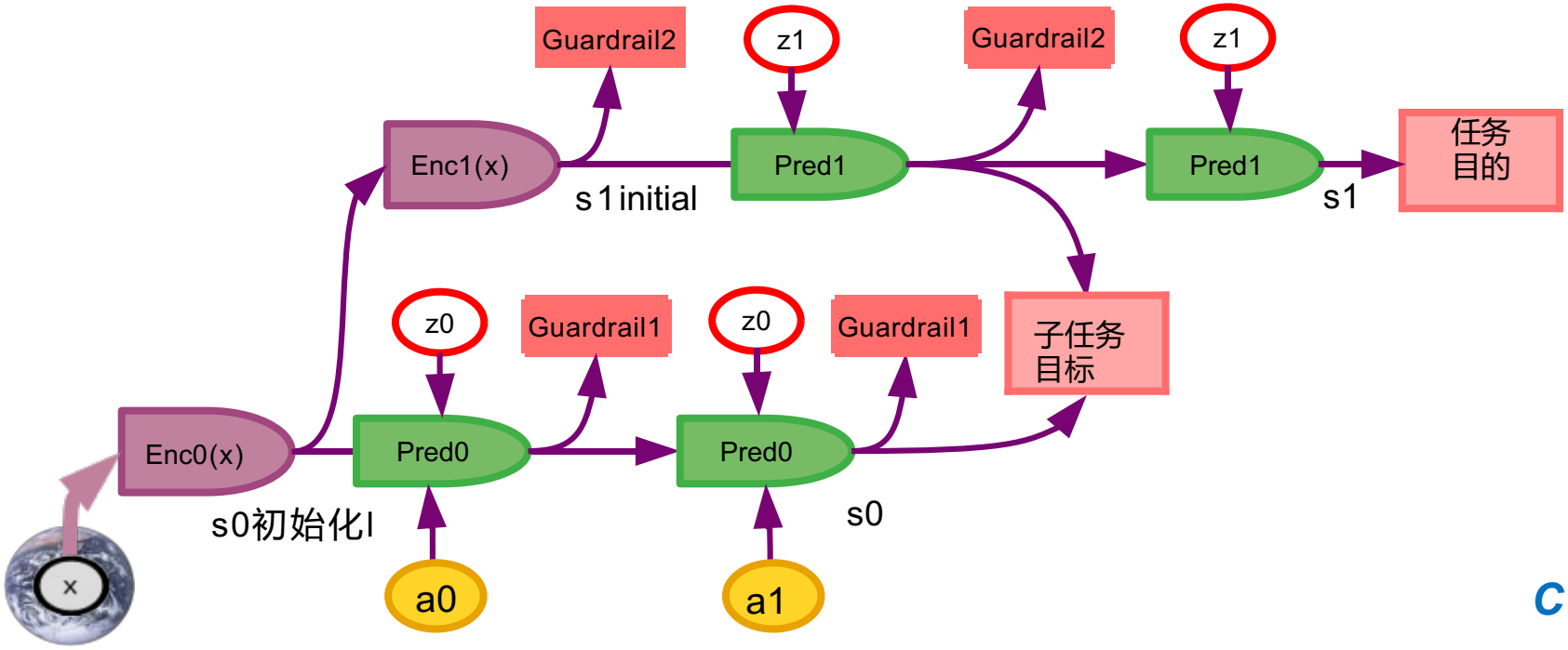
目标驱动的人工智能：非确定性世界模型

- **世界不是确定的或完全可预测的**
- **潜在变量参数化了一组合理的预测**
  - 可以从先前的样本中采样，也可以通过集合进行扫描。
  - 可以针对最坏情况或平均情况进行规划
  - 结果的不确定性可以预测和量化



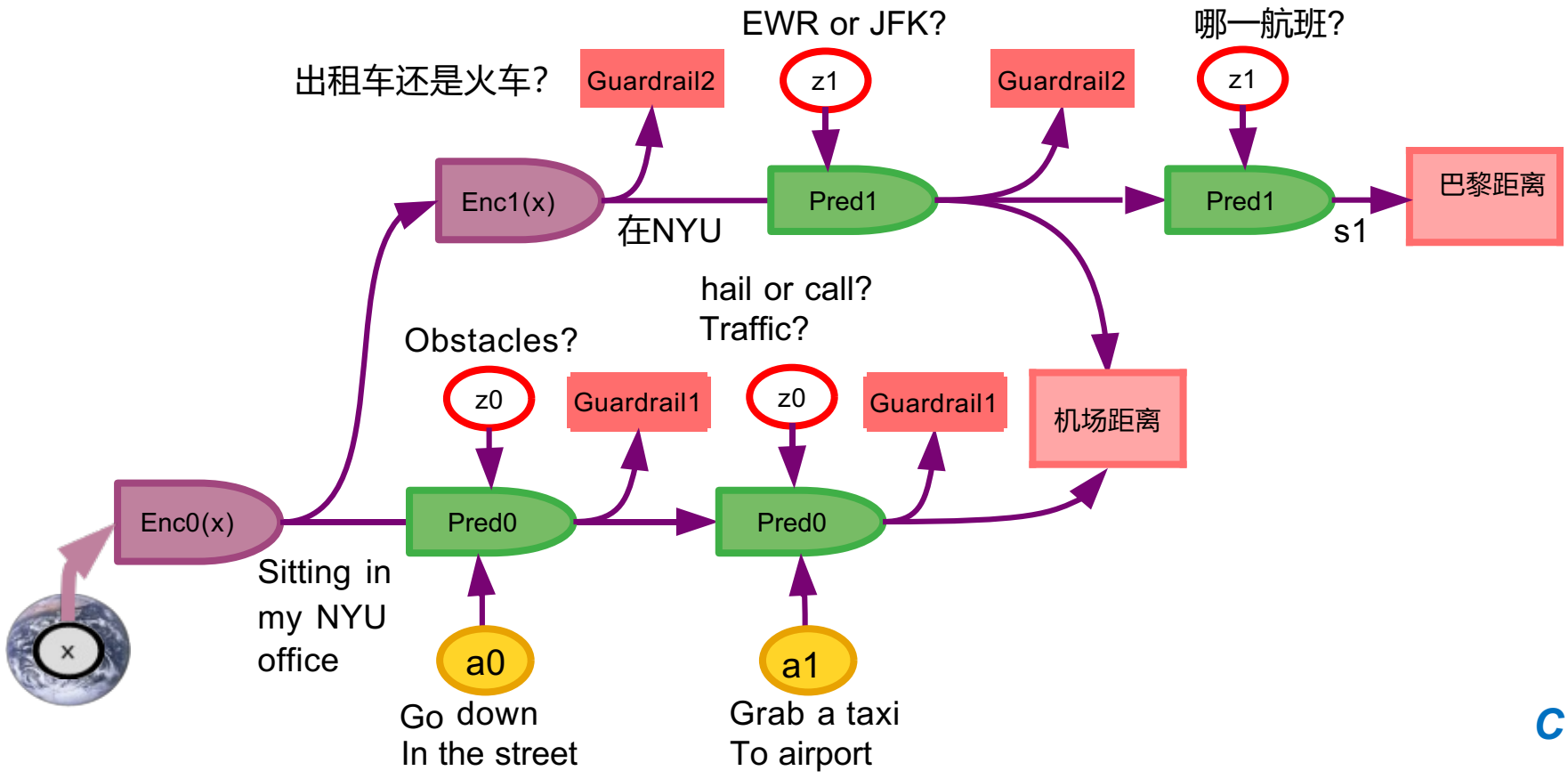
目标驱动的 AI：分层规划

- 分层世界模型与规划
  - 较高级别以更抽象的表示形式进行长期预测
  - 较高级别的预测状态定义较低级别的子任务目标
  - 护栏物镜确保各层安全



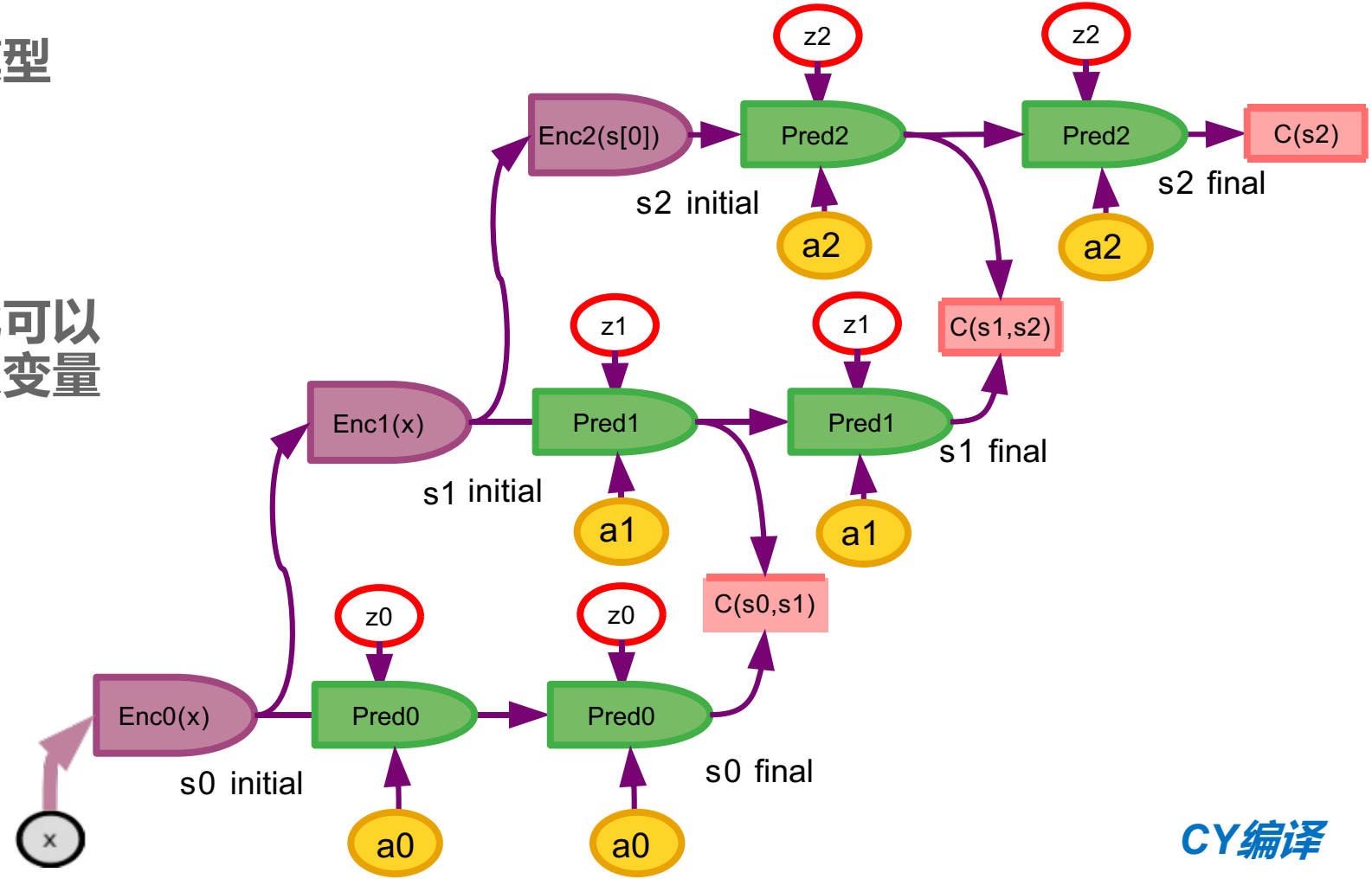
目标驱动的 AI：分层规划

分层规划：从纽约大学到巴黎



目标驱动的 AI：分层规划

- 多层次的世界模型
- k级的预测状态
  确定子任务
  对于 K-1 级
- 基于梯度的优化可以
  应用于操作各级变量
- 可以应用采样
  到潜在变量
  在各个层面。

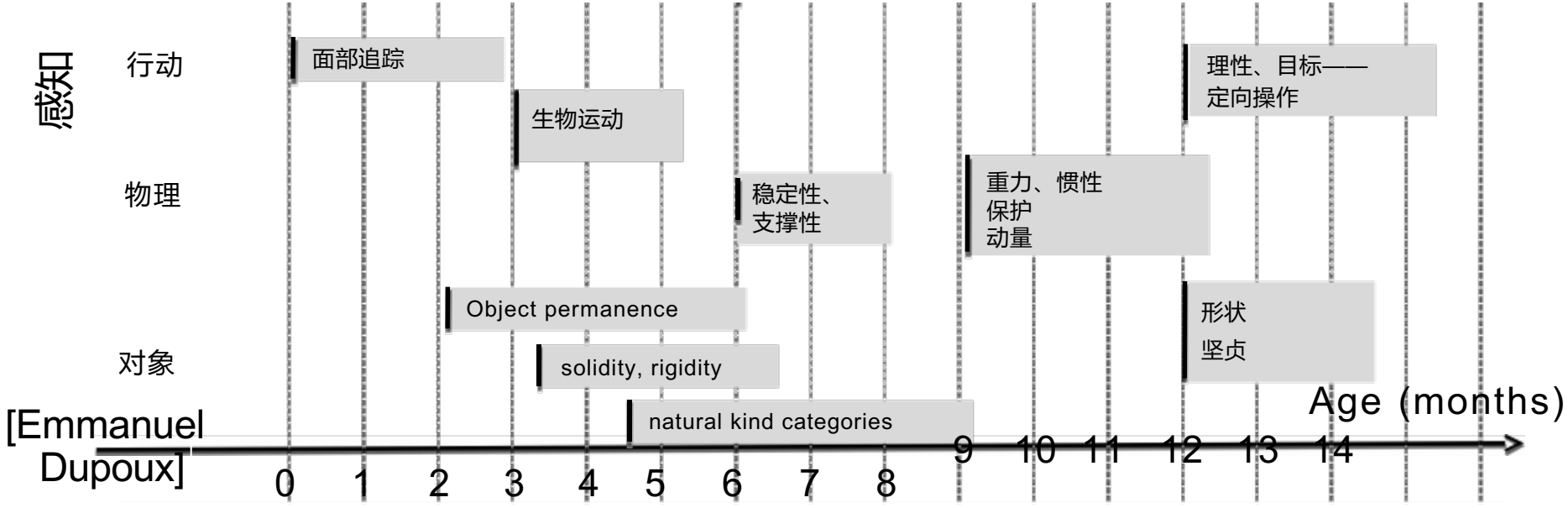


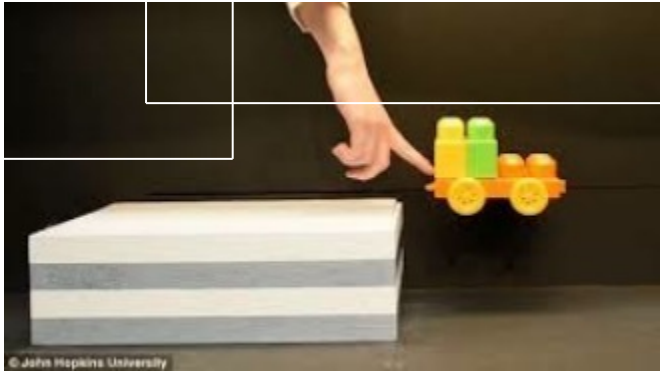
机器怎么可能
学习世界模型
来自感官输入？

跟

自我监督学习

机器如何像动物和人类一样学习?



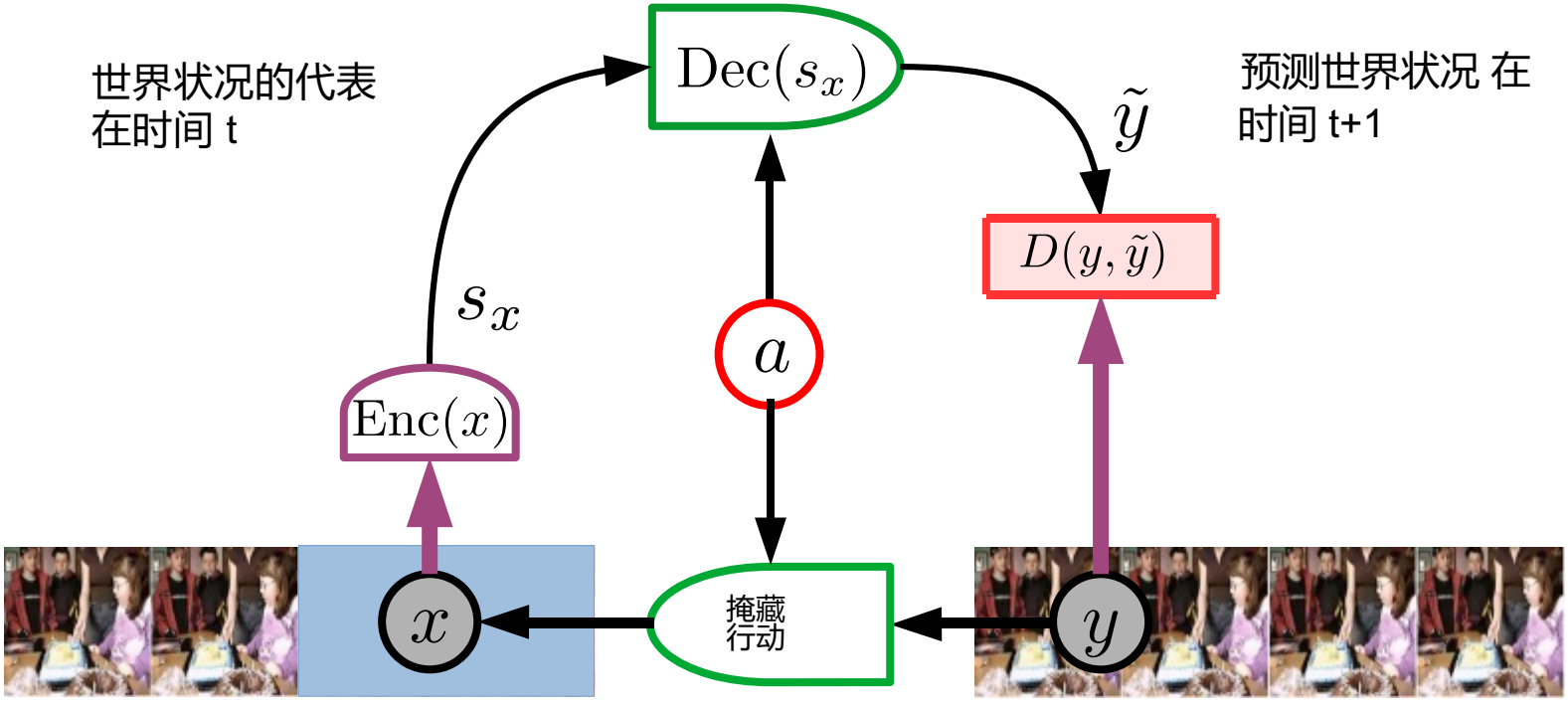


- 宝宝如何学习
- 世界如何
- 工程?

CY编译

具有自我监督训练的生成式世界模型?

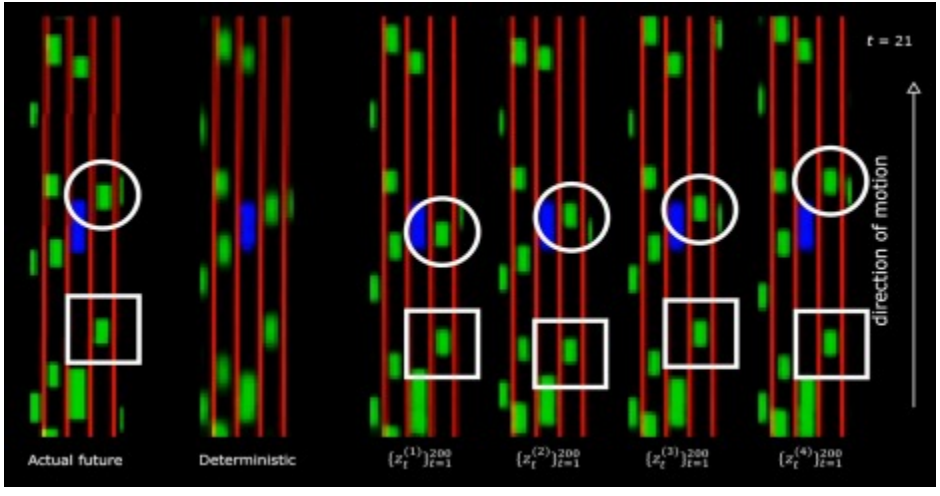
生成式世界模型架构



衍生式架构不适用于图像

- 因为世界只是部分可预测的
- 预测模型应
  表示多个预测
- 概率模型是
  在高暗度连续域中难以处理。
- 生成模型必须预测世界的每一个细节
- My solution: Joint-Embedding Predictive Architecture
  我的解决方案：联合嵌入预测架构

[Mathieu, Couprie, LeCun ICLR 2016]

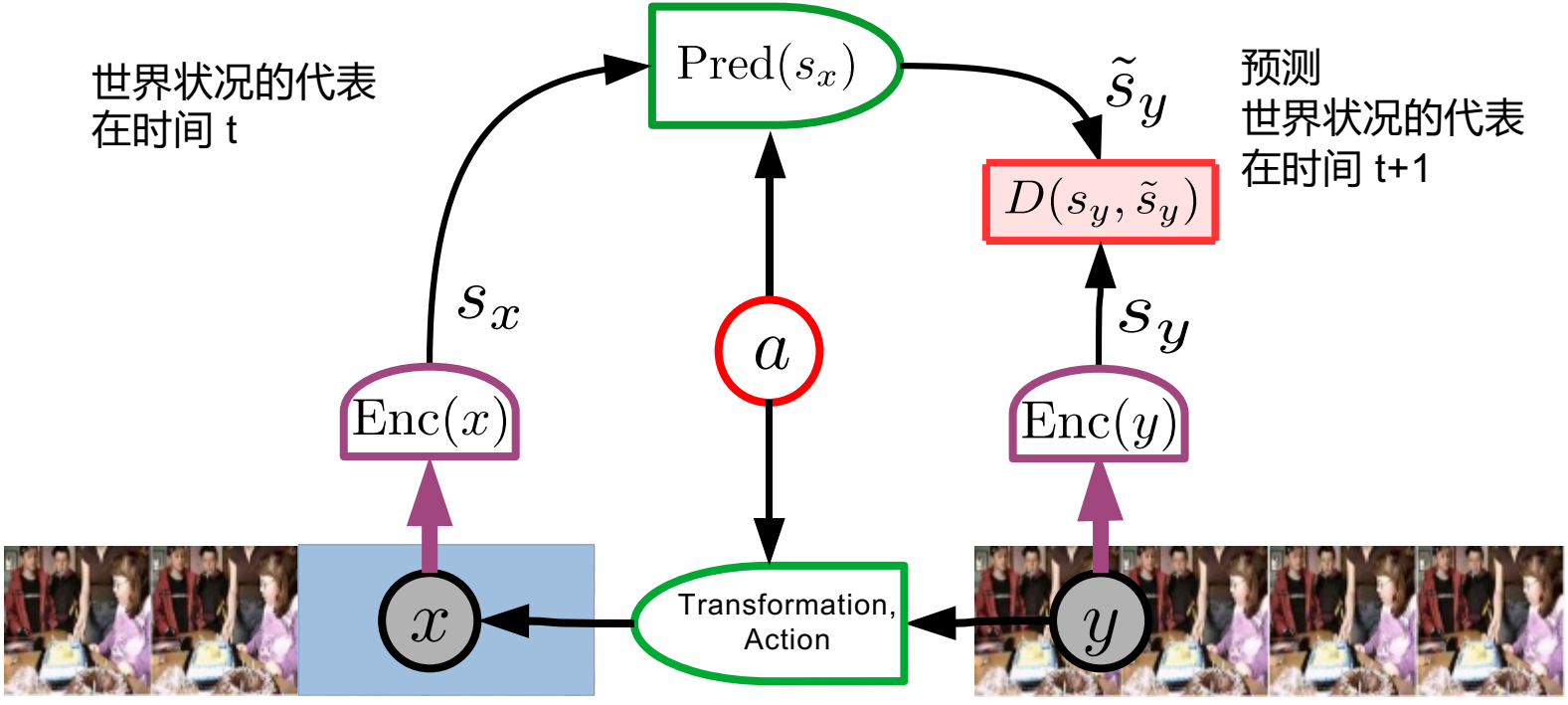


[Henaff, Canziani, LeCun ICLR 2019]

CY编译

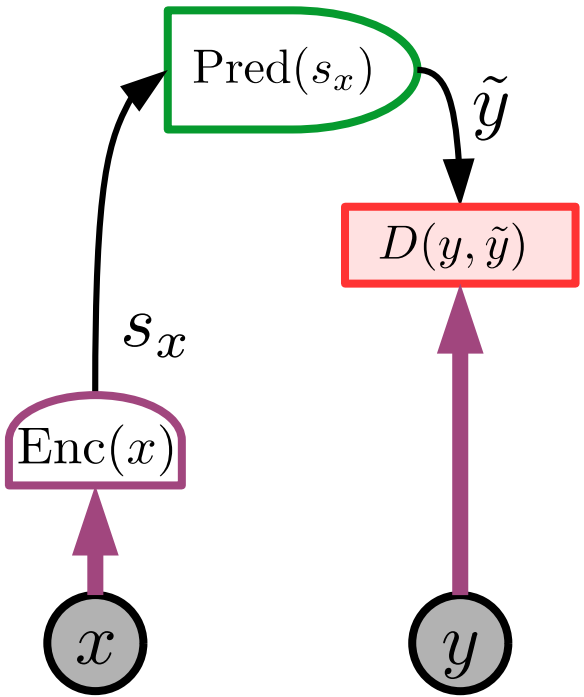
联合嵌入世界模型：自我监督训练

- **联合嵌入预测架构[LeCun 2022], [Assran 2023]**

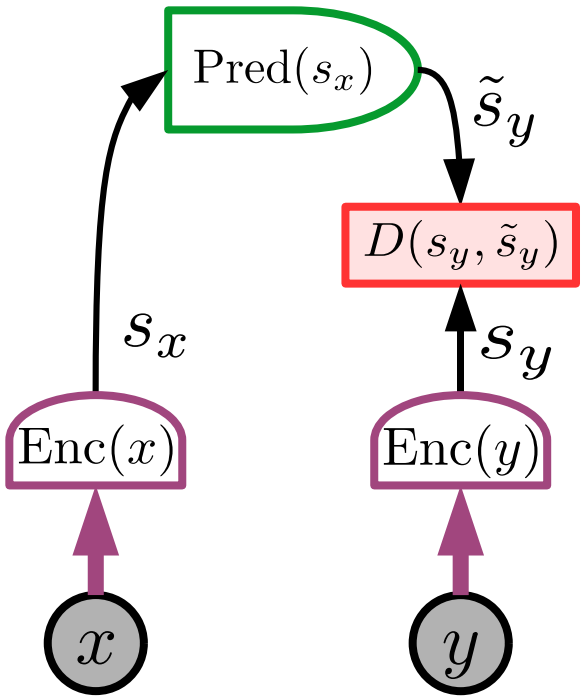


架构：生成式嵌入与联合嵌入

- 生成：预测 y（包含所有细节，包括不相关的细节）
- 联合嵌入：预测 y 的抽象表示



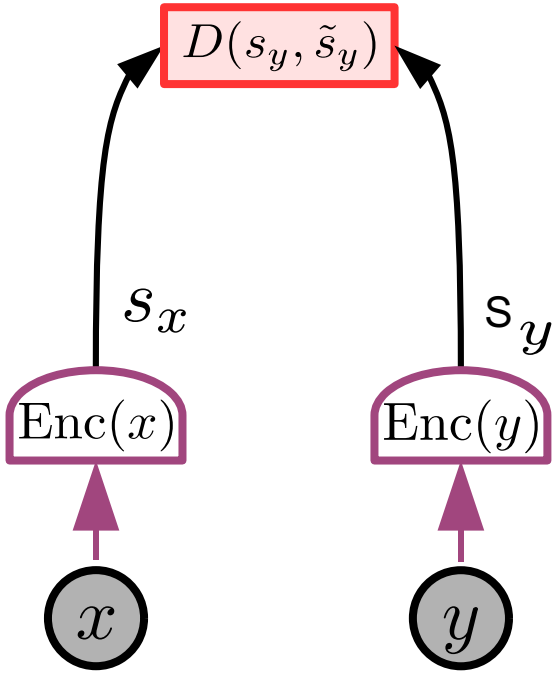
a)衍生式架构
示例：VAE、MAE...



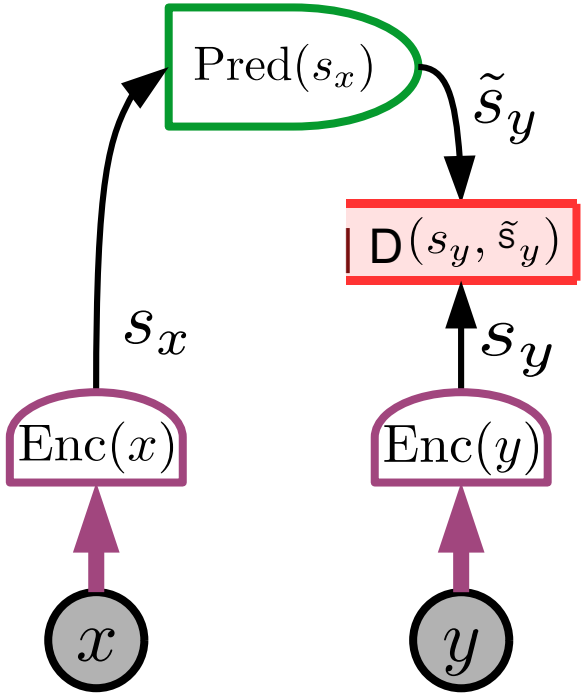
b） 联合嵌入架构

联合嵌入架构

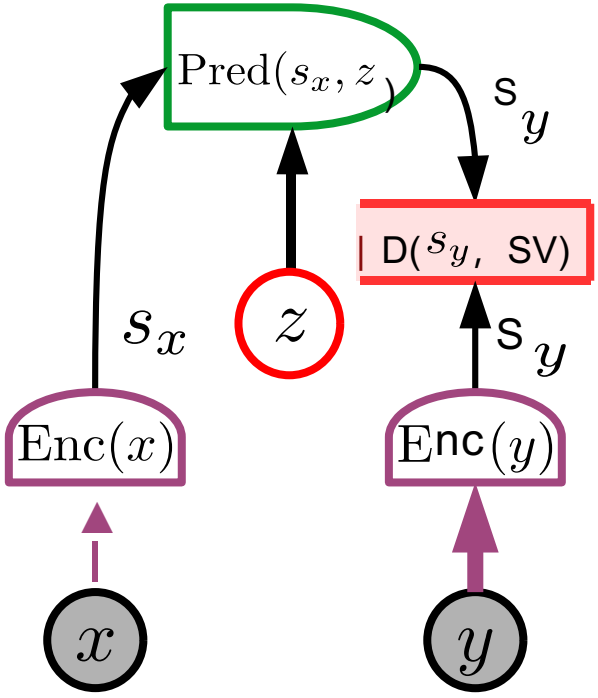
- 计算 x 和 y 的抽象表示
- 尝试使它们彼此相等或可预测。



a） 联合嵌入架构 （JEA） 示例：
Siamese Net、Pirl、MoCo、SimCLR
、BarlowTwins、VICReg、



b)确定性联合嵌入预测架构 （
DJEPA）
示例：BYOL、VICRegL、I-JEPA

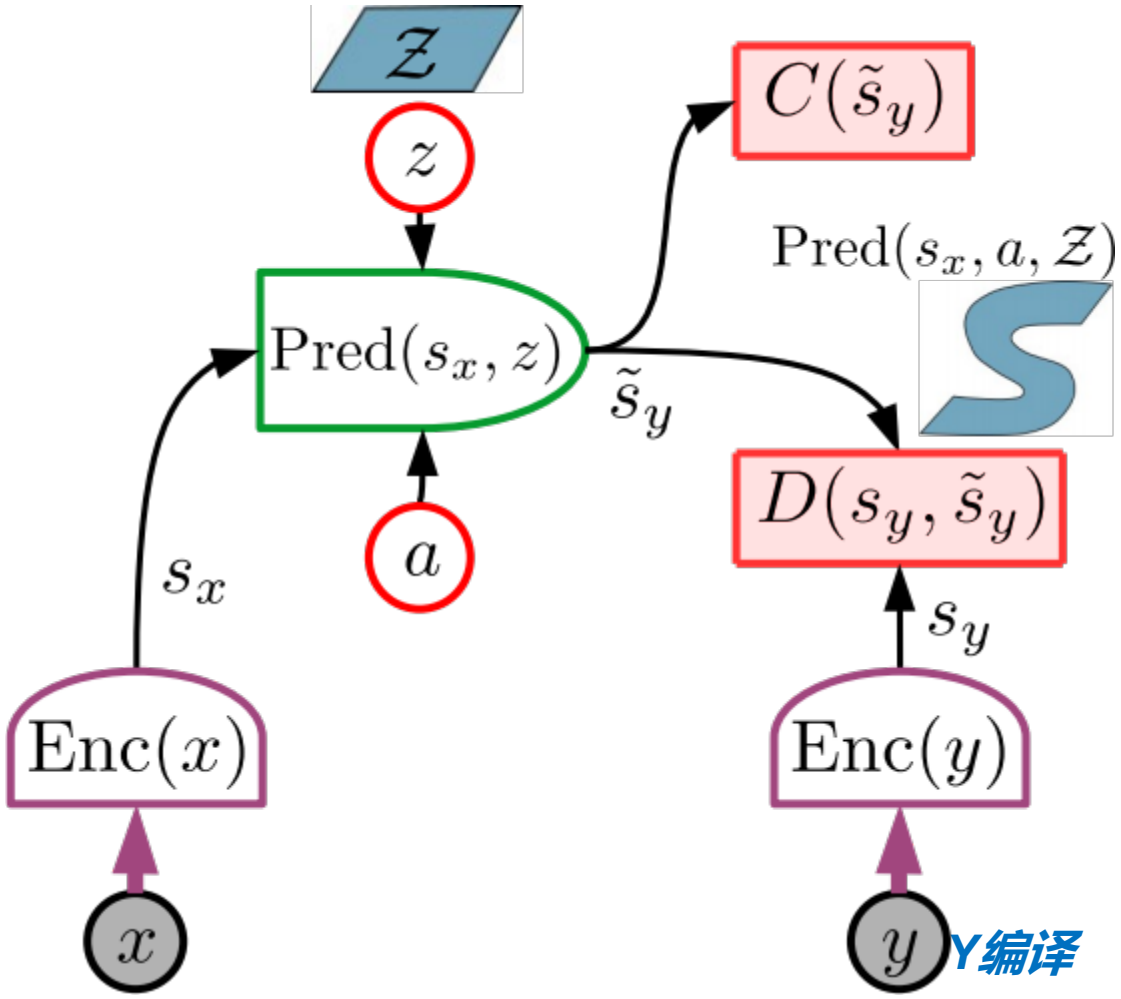


c)联合嵌入预测架构 （JEPA）
示例：等变 VICReg I-JEPA ... ..

CY编译

世界架构模型：JEPA

- **JEPA：联合嵌入预测架构.**
  - x:观察过去和现在
  - y:未来
  - a:行动
  - z:潜在变量（未知）
  - D( ):预测成本
  - C( ):代理成本
  - JEPA预测未来 S y 的表示 过去和现在的表示 S x

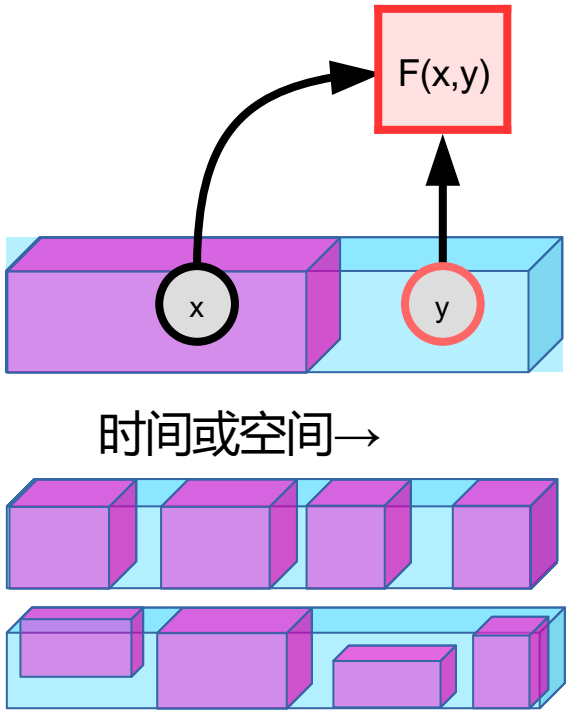


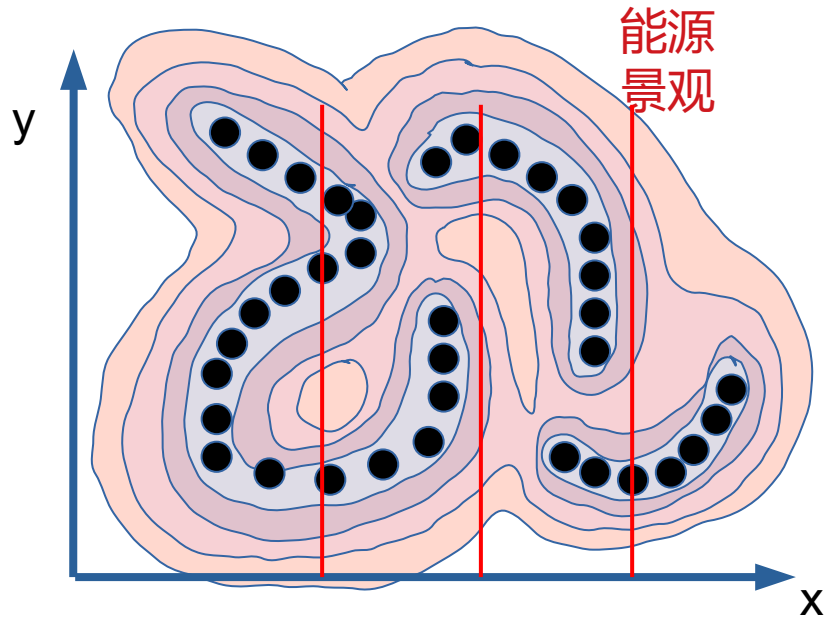
基于能量的模型

通过能量函数捕获依赖关系

基于能量的模型：隐式函数

- **形式化和理解所有模型类型的唯一方法**
  - 为兼容的 x 和 y 对提供低能量
  - 为不相容的对提供更高的能量





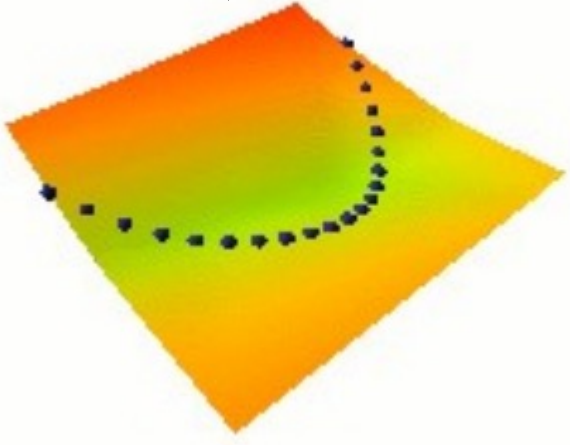
$$\check{y} = \text{argmin}_y F(x, y)$$

CY编译

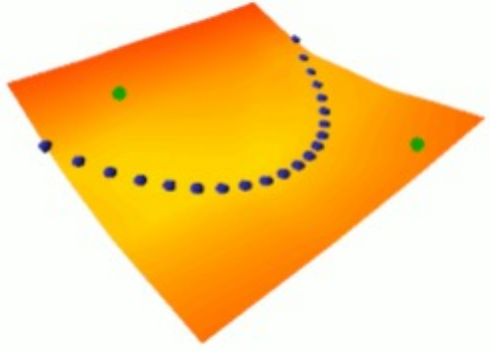
训练基于能量的模型：防止倒塌

- 训练基于弹性能量的模型
- **我们需要一个损失函数来塑造能量面，以便：**
  - 数据点能量低
  - 高数据密度区域之外的点具有更高的能量.

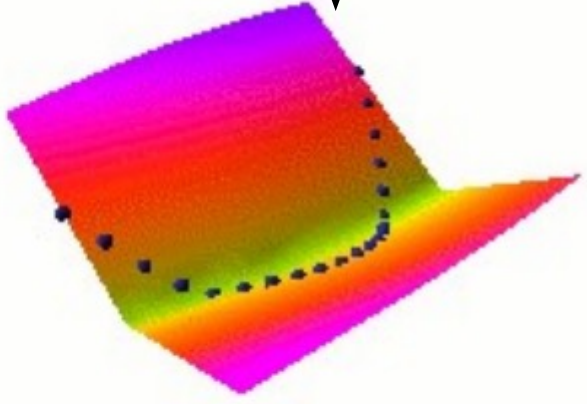
崩溃!



对比法

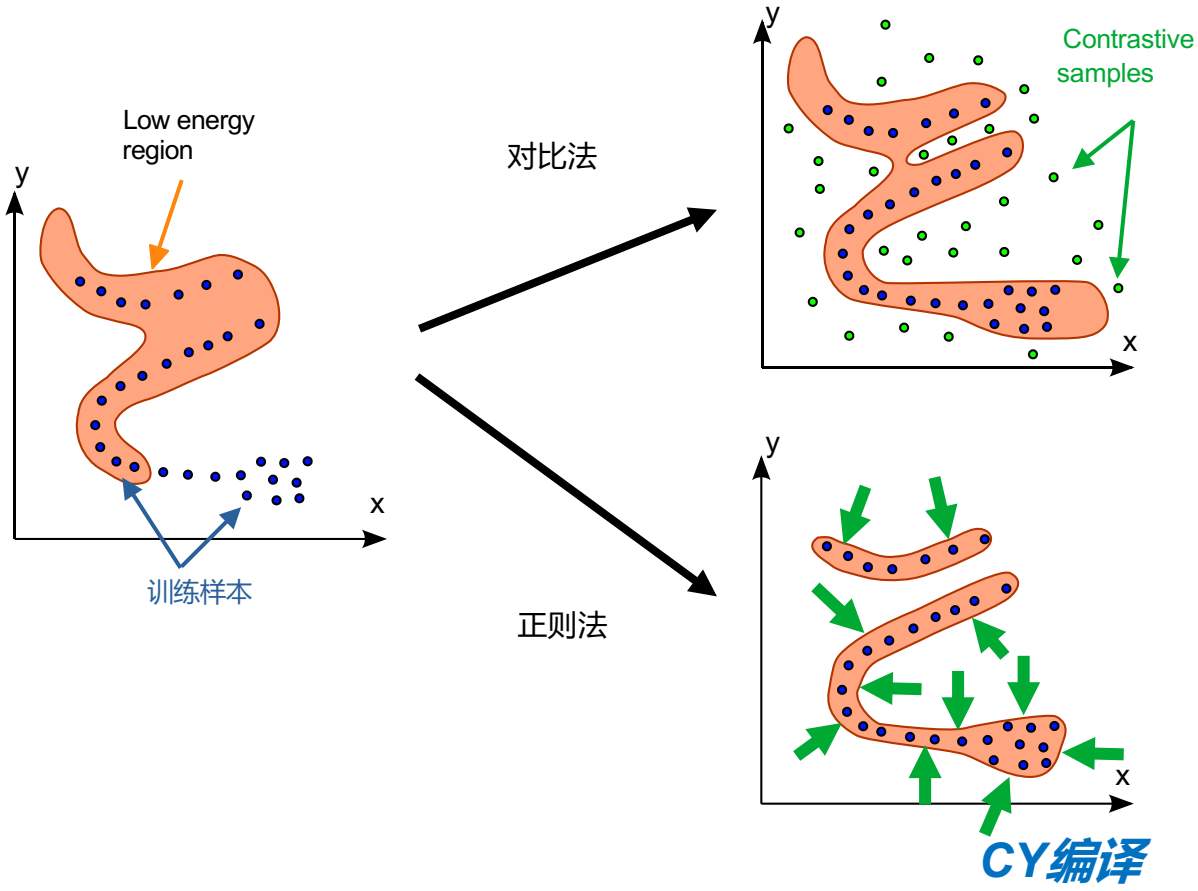


正则化方法



EBM训练：两类方法

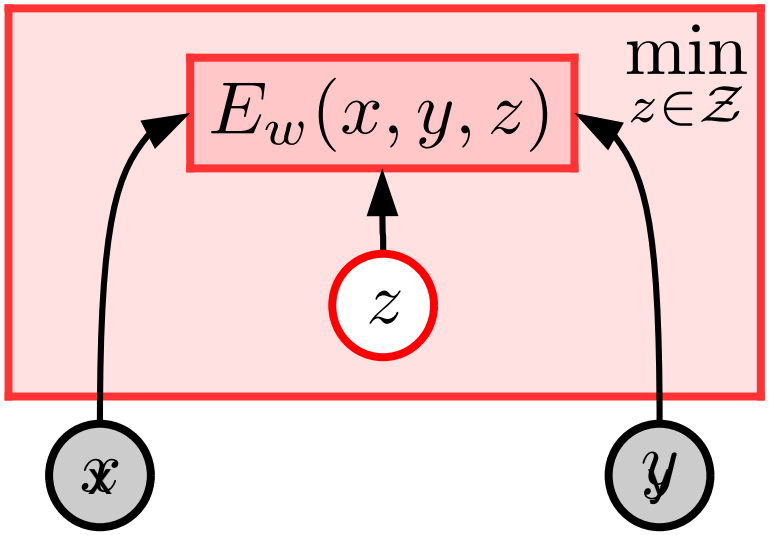
- **对比方法**
  - 降低训练样本的能量
  - 拉起能量 适当生成对比样品
  - 尺寸缩放非常糟糕
- **正则化方法**
  - 正则化器最大限度地减少了可以消耗低能量的空间体积



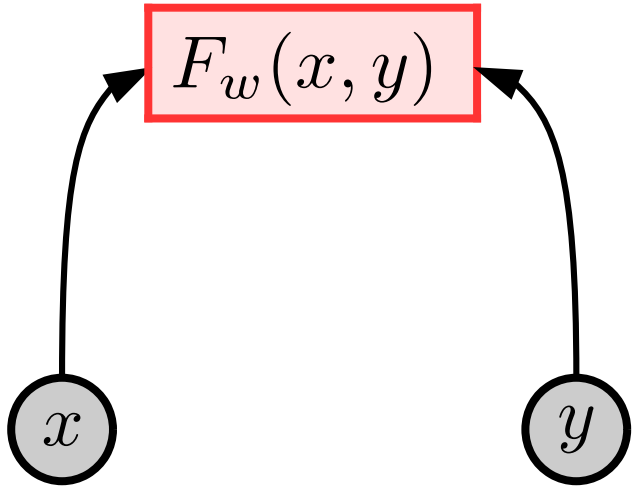
潜在可变 EBM

- **潜在变量z:**
  - 捕获 y 中不可用的信息x
  - 通过最小化计算

$$\check{z} = \text{argmin}_{z \in \mathcal{Z}} E_w(x, y, z) \qquad F_w(x, y) = E_w(x, y, \check{z})$$

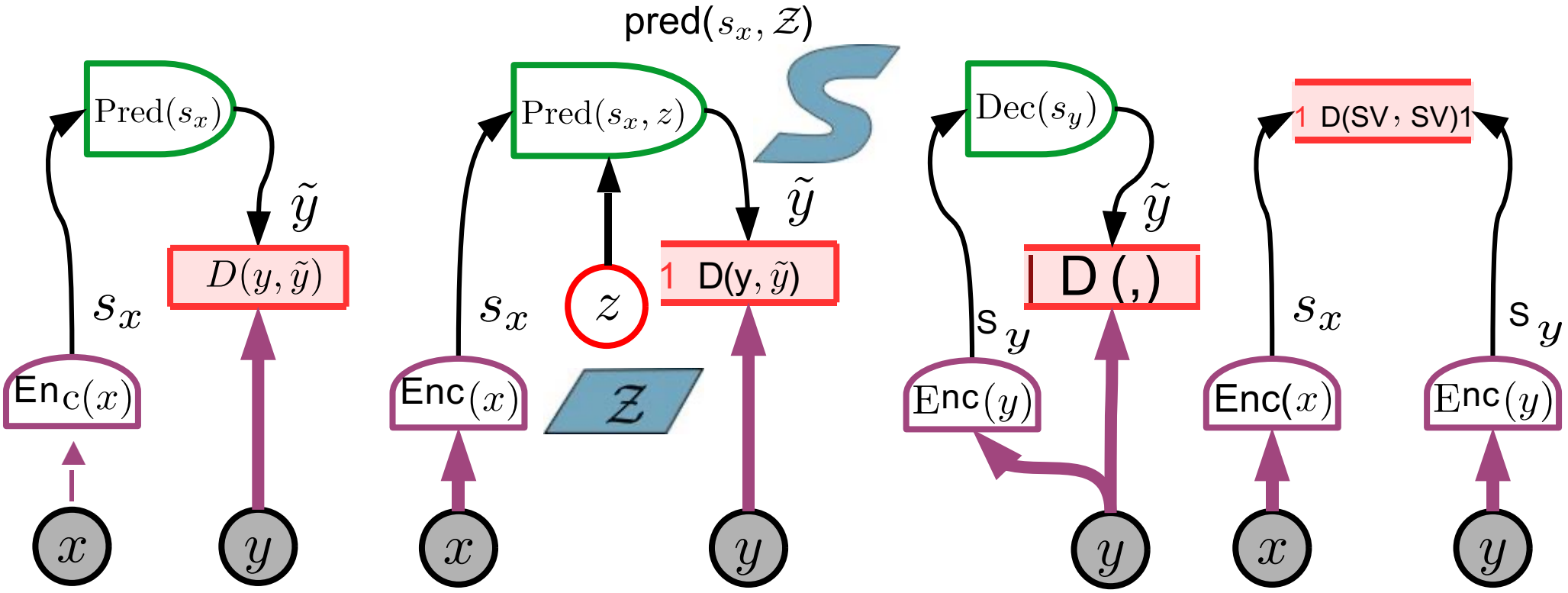


=



EBM 架构

- 一些架构可能导致能量面的坍塌



a) Prediction / regression
NO COLLAPSE

b) Generative latent-variable Architecture
CAN COLLAPSE

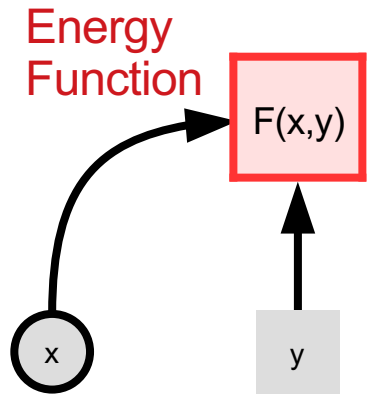
c) Auto-Encoder
CAN COLLAPSE

d) Joint Embedding Architecture
CAN COLLAPSE

基于能量的模型与概率模型

- **概率模型是EBM的一个特例**
  - 能量就像未归一化的负对数概率
- **为什么使用 EBM 而不是概率模型?**
  - 择评分功能方面具有更大的灵活性。
  - 灵活选择学习目标函数
- **从能量到概率：吉布斯-玻尔兹曼分布**
  - Beta 是一个正常数

EBM在选
More



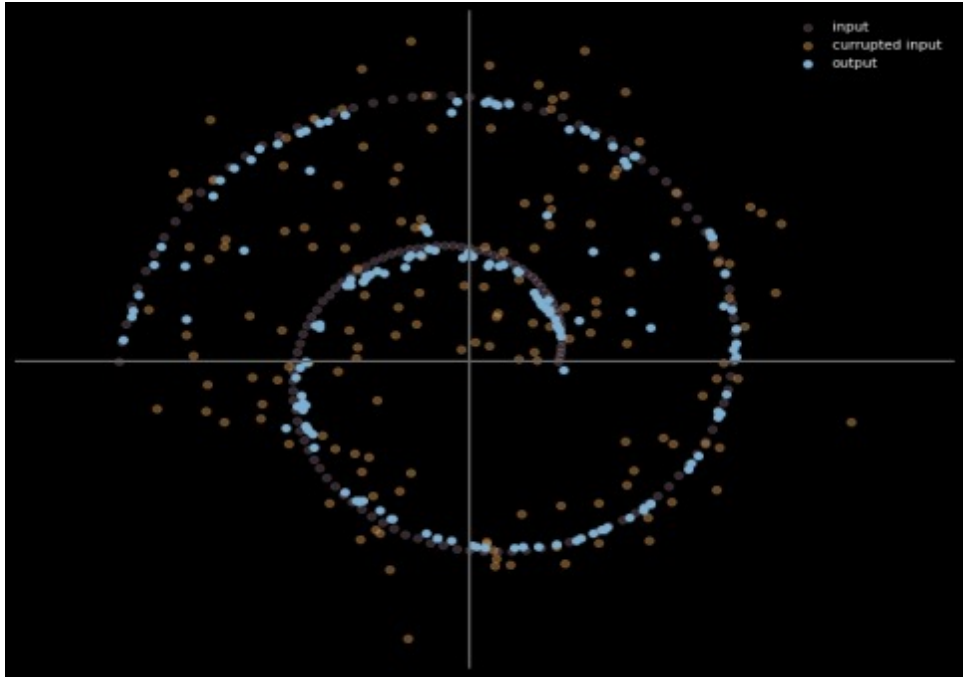
$$P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y')}}$$

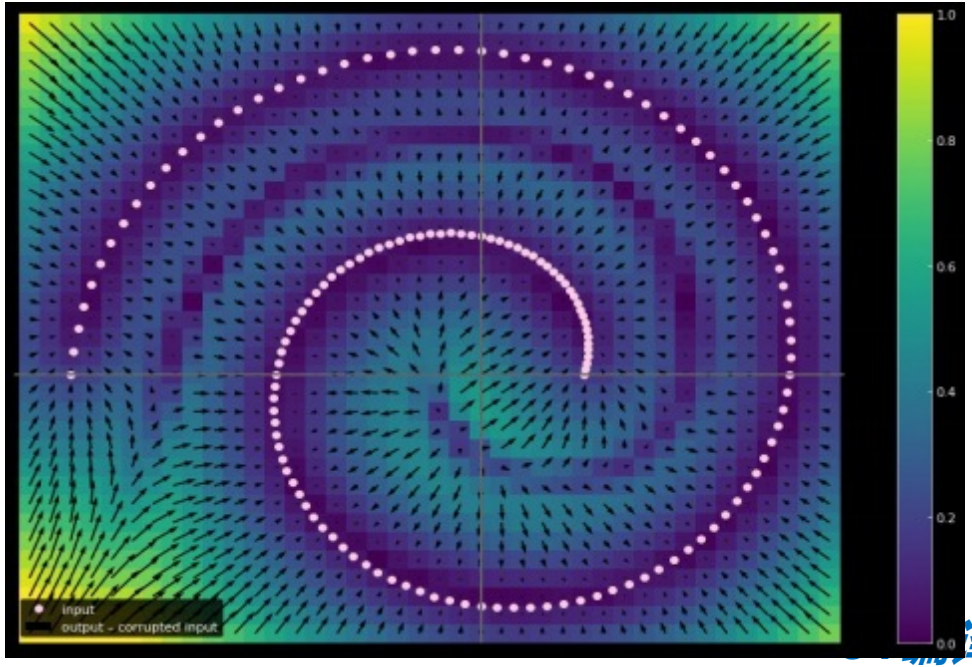
对比方法与正则化/架构方法

- **对比：[它们都是选择向上推点的不同方法]**
- C1:下推数据点的能量，在其他任何地方上推：最大似然（需要可处理的分区函数或变分近似）
- C2:下推数据点的能量，上推选定位置：MC/MMC/HMC 的最大似然、对比散度、度量学习/连体网、比率匹配、噪声
- 对比估计、最小概率流、对抗生成器/GAN
- C3:训练将数据流形上的点映射到数据流形上的点的函数：降噪自动编码器、屏蔽自动编码器(e.g. BERT)
- A2**正则化/架构：[限制潜在表示信息容量的不同方法]**
- **A1：构建机器，使低能空间的体积是有界的：PCA、K-means、高斯混合模型、平方 ICA、归一化流动……**: 使用正则化项来测量具有低能量的空间体积：稀疏编码、稀疏自动编码器、LISTA、变分自动编码器、离散化/VQ/VQVAE。
- A3: F(x,y) = C(y, G(x,y)), make G(x,y)相对于 y 尽可能“恒定”：收缩自动编码器，饱和自动编码器
- A4:最小化梯度并最大化数据点周围的曲率：分数匹配

对比 EBM 培训：去噪自动编码器

- [LeCun 1987], [Seung 1998], [Vincent 2008, 2010]
- NLP: BERT [



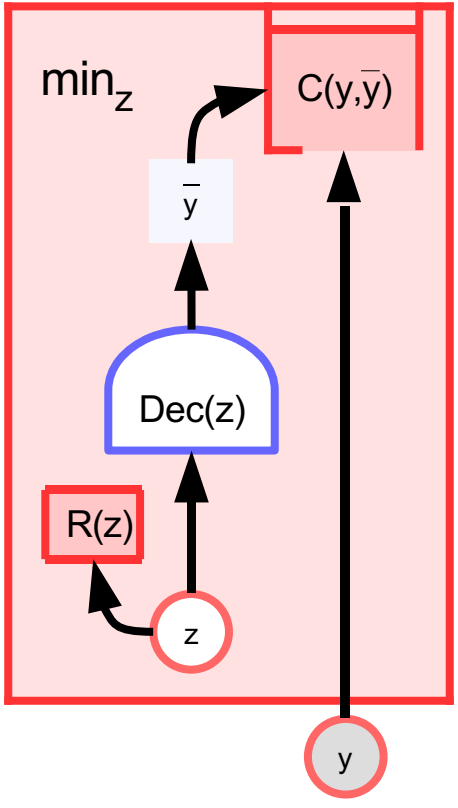


Figures: Alfredo Canziani

示例：正则化潜在变量 EBM

- **基本思想：限制潜在变量的信息容量，限制低能量区域的体积**

- **示例：K-Means、稀疏编码**



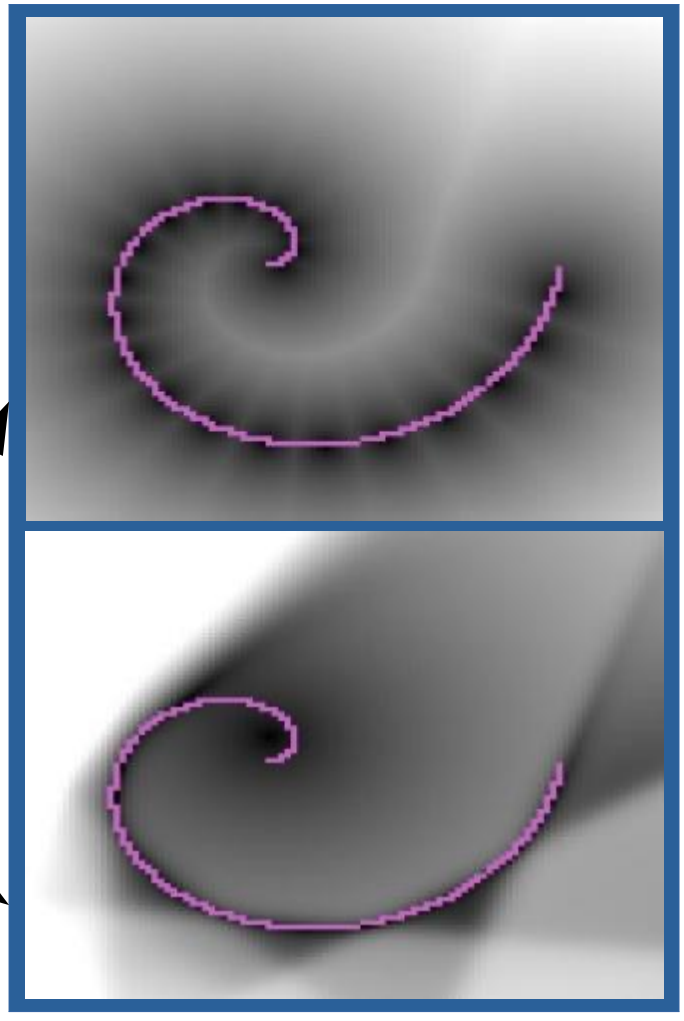
$$E(y,z) = C(y, \mathrm{Dec}(z)) + R(z)$$

$$F_\infty(x,y) = \min_z E(x,y,z)$$

$$\check{y}, \check{z} = \mathrm{argmin}_{y,z} E(x,y,z)$$

$$E(y,z) = ||y - Wz||^2 \quad z \in 1hot$$

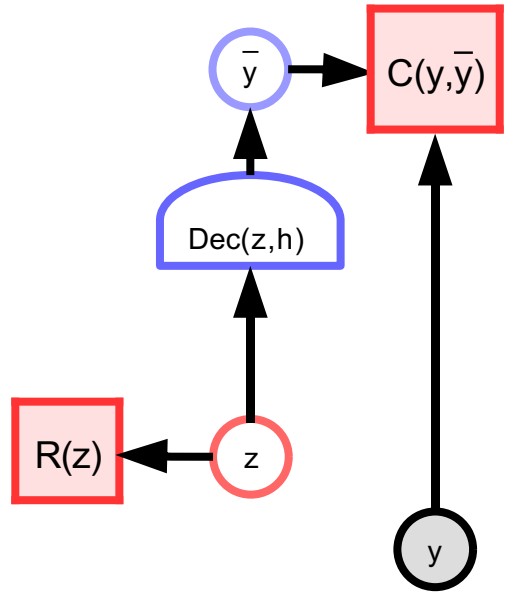
$$E(y,z) = ||y - wz||^2 + \lambda |z|_{L1|}$$



通过使 z “模糊”（随机）来正则化 z

- **必须最小化潜在变量 z 的信息内容**
- **一种（概率）方法可以做到这一点：**
  - make z “fuzzy” (e.g. stochastic)
  - z是分布中的样本q(z|y)

$$E(y,z) = C(y, Dec(z))$$



- **最小化下能量的期望值q(z|y)**

$$< E(y) >= \int_z q(z|y) E(y,z)$$

- **最小化 q（z|y） 中关于 y 的信息含量**

最小化 z 的预期能量和信息含量

- **将预期能量降至最低**

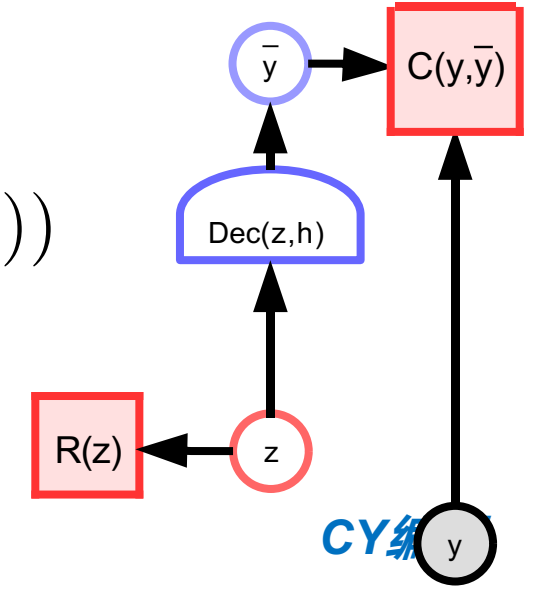
$$< E(y) >_q = \int_z q(z|y) E(y,z)$$

$$E(y,z) = C(y, Dec(z))$$

- **最小化相对熵**
  - Between q(z|y) and a prior distribution p(z).

$$KL(q(z|y), p(z)) = \int_z q(z|y) \log_2 (q(z|y)/p(z))$$

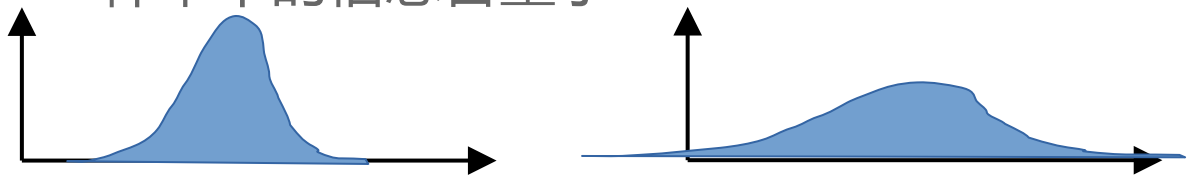
- 这是 q（z|y） 中一个样本的位数
  会给我们关于 z 的信息，知道 z 来自 p（z）



变分自由能：以 z 为单位交易平均能量和信息

- **找到一个分布 q（z|y），该分布在最小化预期能量的同时具有最大熵**

- 高熵分布 == 样本中的信息含量小



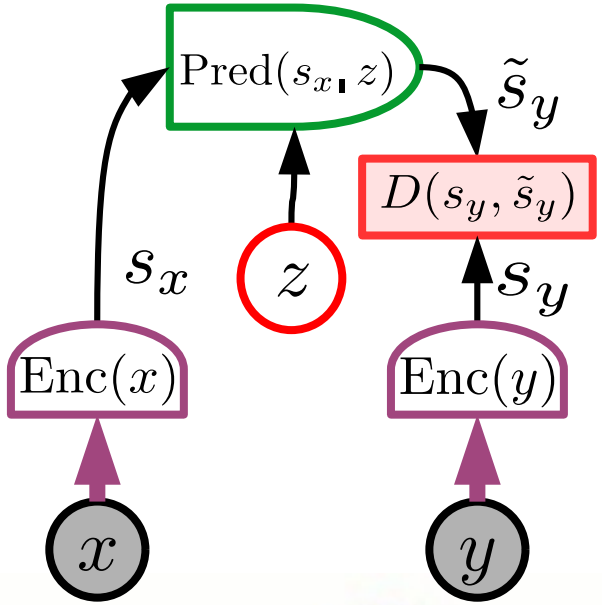
- 选择一个分布族 q（z|y）（例如高斯分布），并找到使变分自由能最小化的分布族:

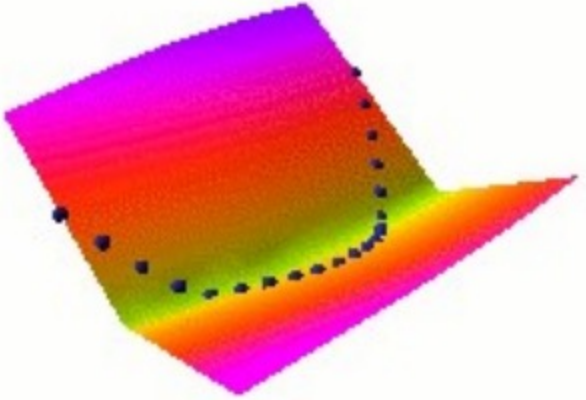
$$\tilde{F}_q(y) = \int_z q(z|y)E(y,z) + \frac{1}{\beta}\int_z q(z|y)\log_2(q(z|y)/p(z))$$

- 能量和熵之间的权衡由 beta 参数控制。

建议:

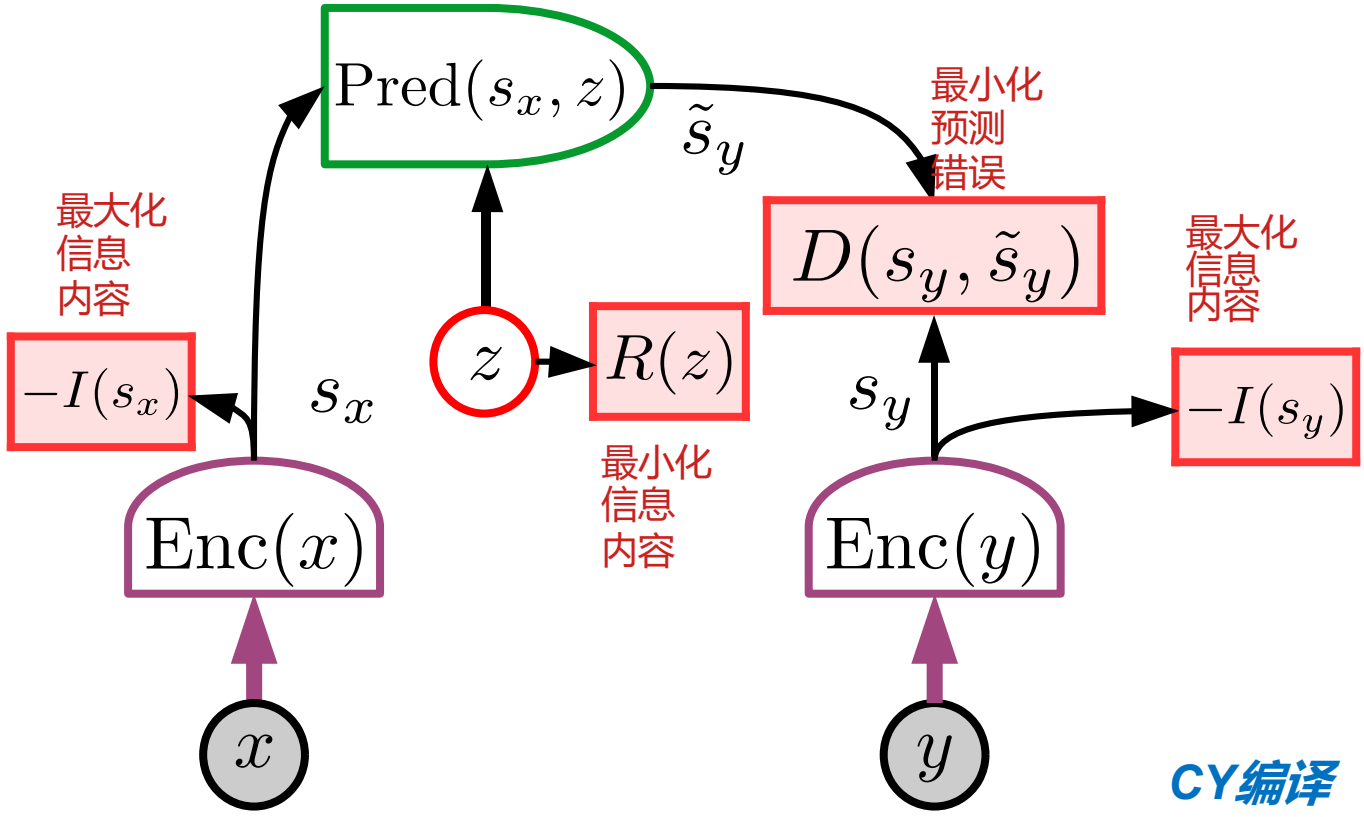
- **放弃生成模型**
  - 支持联合嵌入架构
- **放弃概率模型**
  - 支持基于能量的模型
- **摒弃对比方法**
  - 赞成正则化方法
- **放弃强化学习**
  - 支持模型预测控制
  - 仅当计划没有产生预测结果时，才使用 RL，以调整世界模型或批评者。





使用正则化方法训练 JEPA

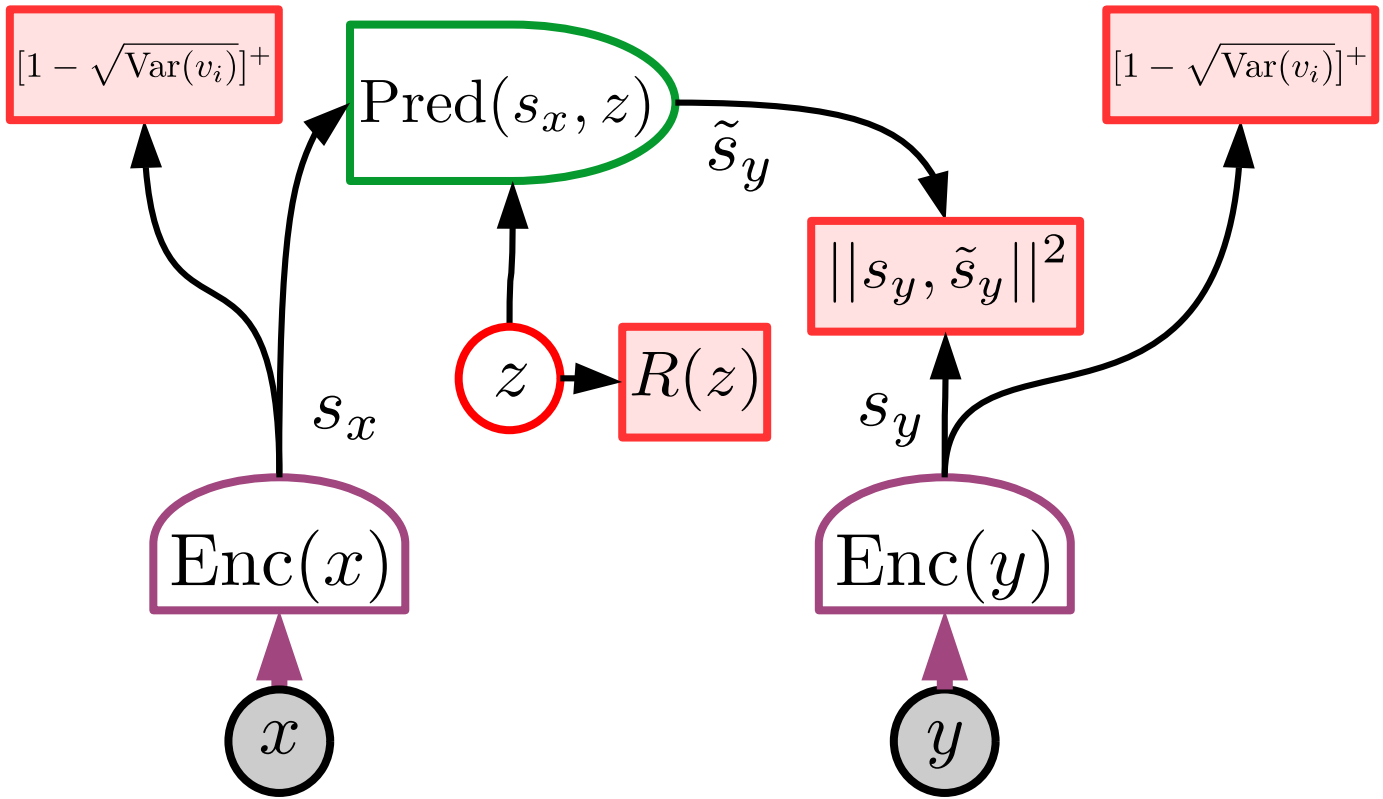
- **成本中的四个术语**
  - 最大化信息内容 x 的表示
  - 最大化信息内容 y 的表示
  - 最小化预测错误
  - 最小化潜在信息含量变量 z



CY编译

VICReg：方差、不变性、协方差正则化

- **方差:**
  - 保持组件的差异交涉
- **不变性:**
  - 最小化预测错误。

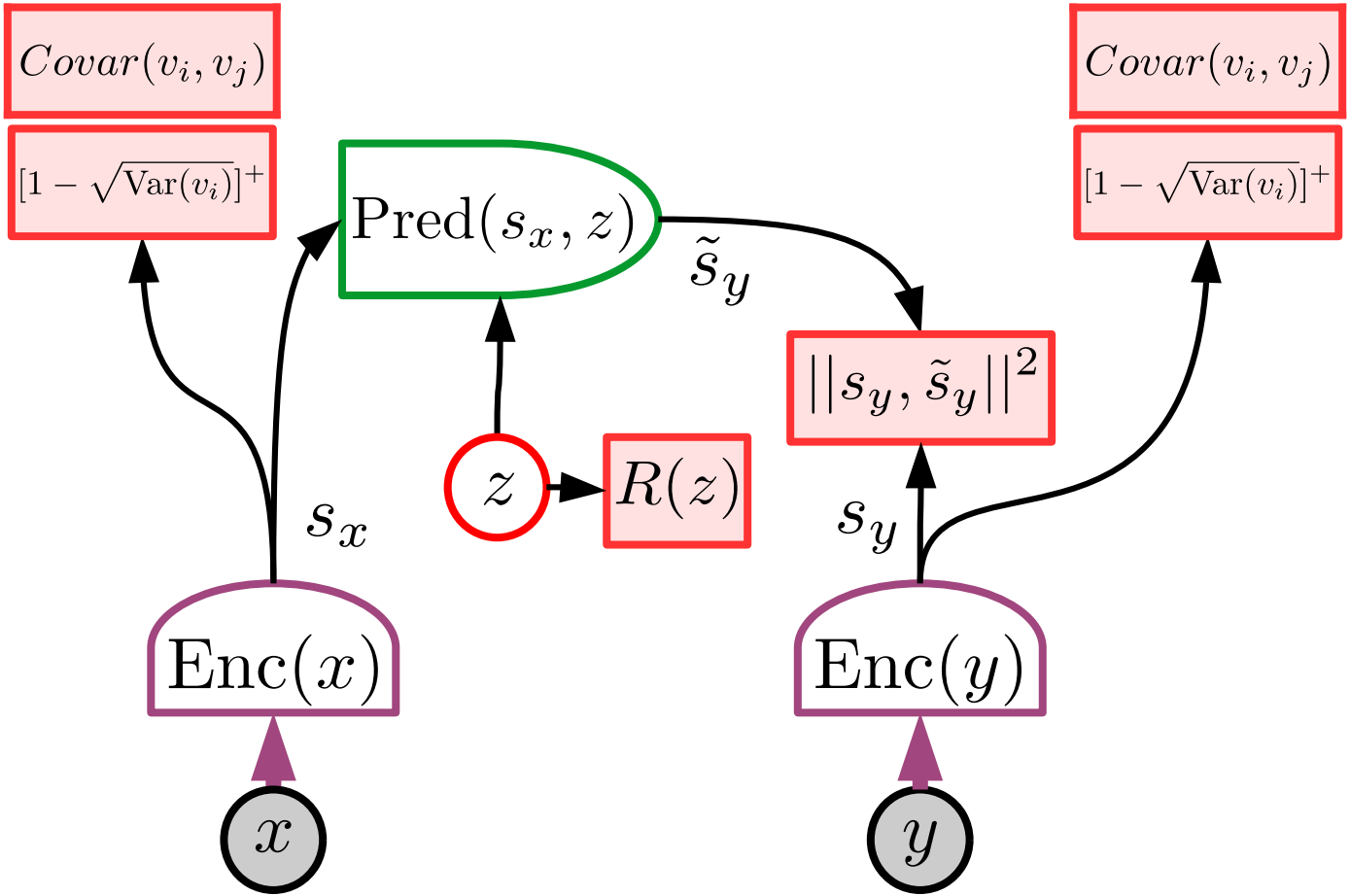


巴洛双胞胎 [Zbontar 等人 ArXiv：2103.03230]、VICReg [Bardes、Ponce、LeCun arXiv：2105.04906、ICLR 2022]、VICRegL [Bardes 等人 NeurIPS 2022]、MCR2 [Yu 等人 NeurIPS 2020][马、曹、沈，2022]

CY编译

VICReg：方差、不变性、协方差正则化

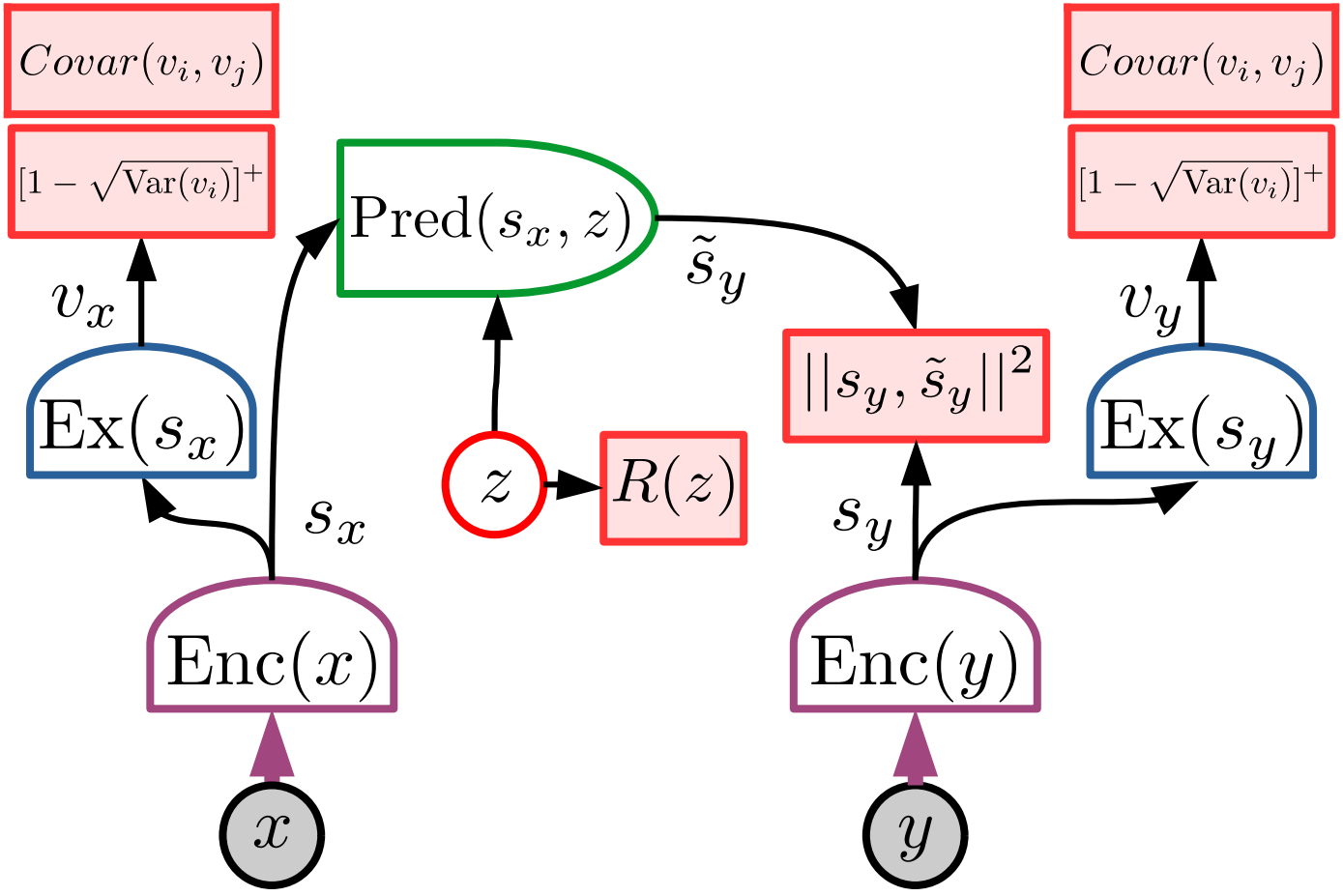
- **方差：**
  - 保持组件的差异交涉
- **协方差：**
  - 去相关的组件
  - 表示的协方差矩阵
- **n方差：**
  - 最小化预测错误。



巴洛双胞胎 [Zbontar 等人 ArXiv：2103.03230]、VICReg [Bardes、Ponce、LeCun arXiv：2105.04906、ICLR 2022]、VICRegL [Bardes 等人 NeurIPS 2022]、MCR2 [Yu 等人 NeurIPS 2020][马、曹、沈，2022]

VICReg：方差、不变性、协方差正则化

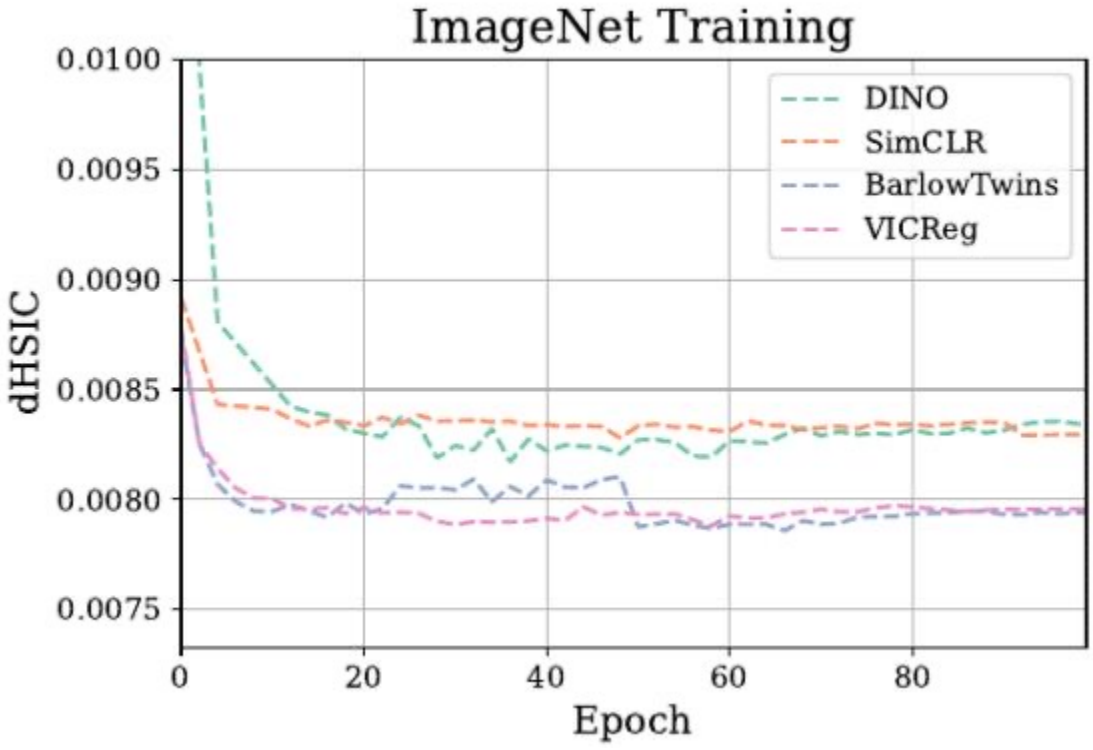
- **方差：**
  - **协方差：**
- 去相关的组件表示的协方差矩阵
- **不变性：**
  - 最小化预测错误。

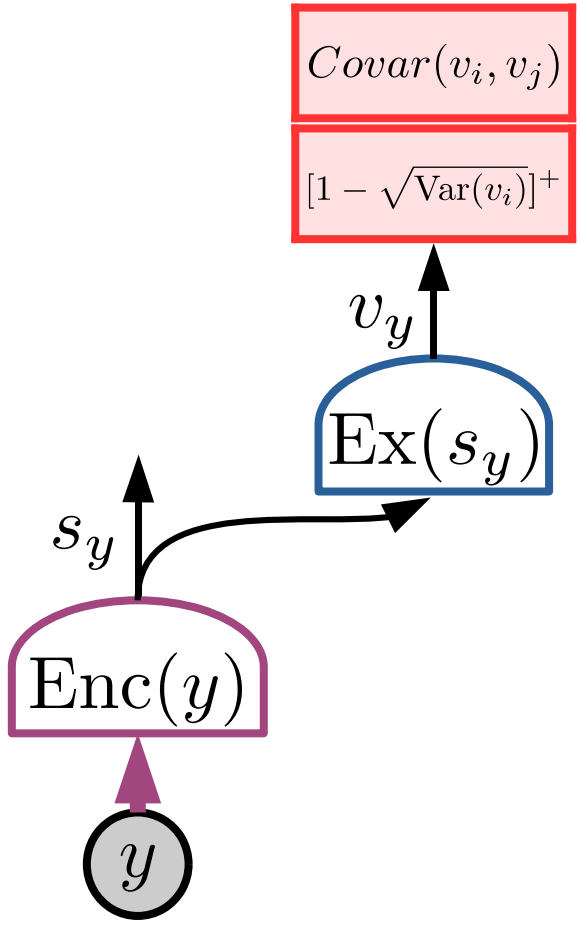


Barlow Twins [Zbontar et al. ArXiv:2103.03230], VICReg [Bardes, Ponce, LeCun arXiv:2105.04906, ICLR 2022], VICRegL [Bardes et al. NeurIPS 2022], MCR2 [Yu et al. NeurIPS 2020][Ma, Tsao, Shum, 2022]

VICReg：扩展器使变量成对独立

- [Mialon, Balestriero, LeCun arxiv:2209.14905]
- VC标准可用于源分离/ICA

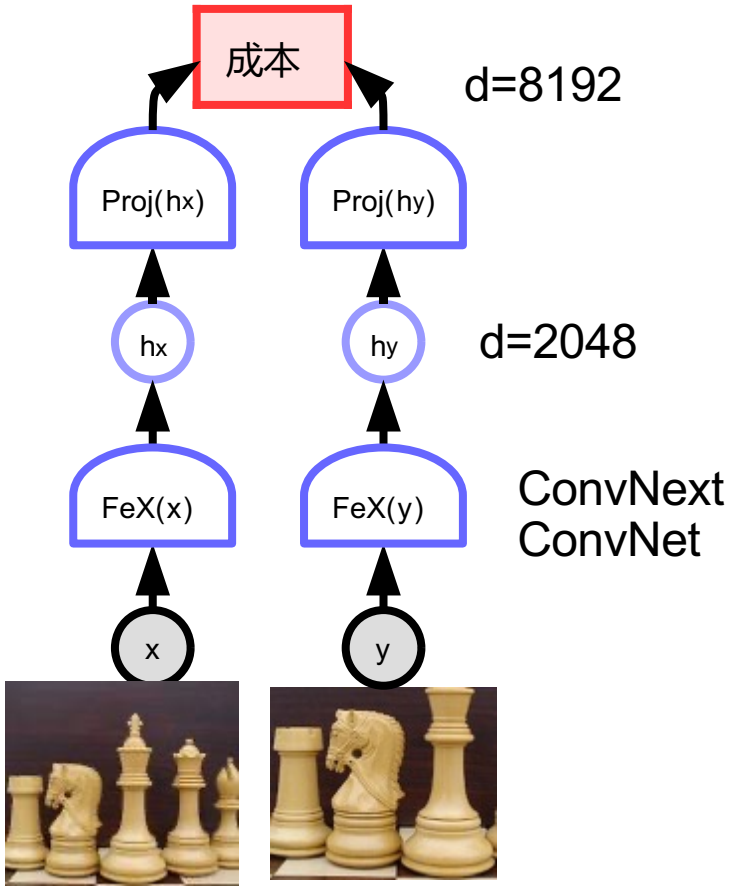




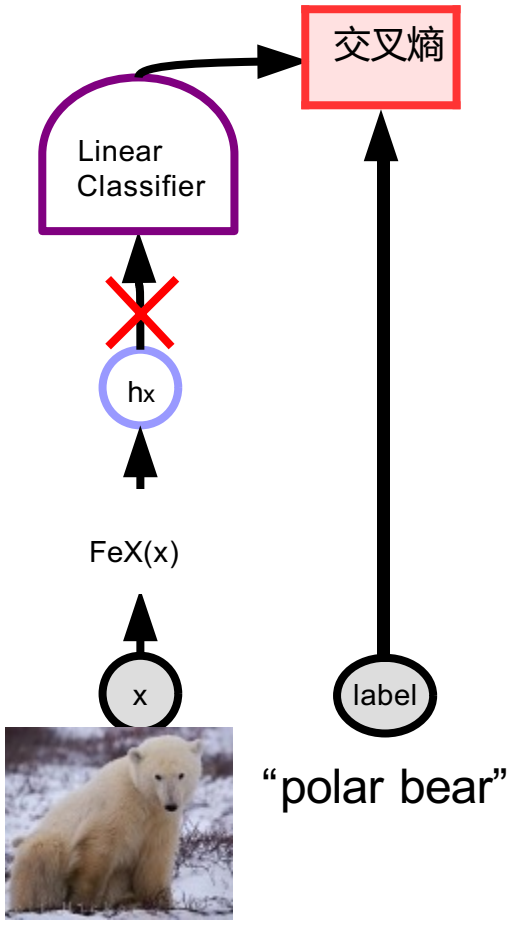
CY编译

用于图像识别的 SSL 预训练联合嵌入

使用 VICReg 进行预训练的 JEA



训练有监督的线性头



CY编译

VICReg：线性头和半监督的结果.

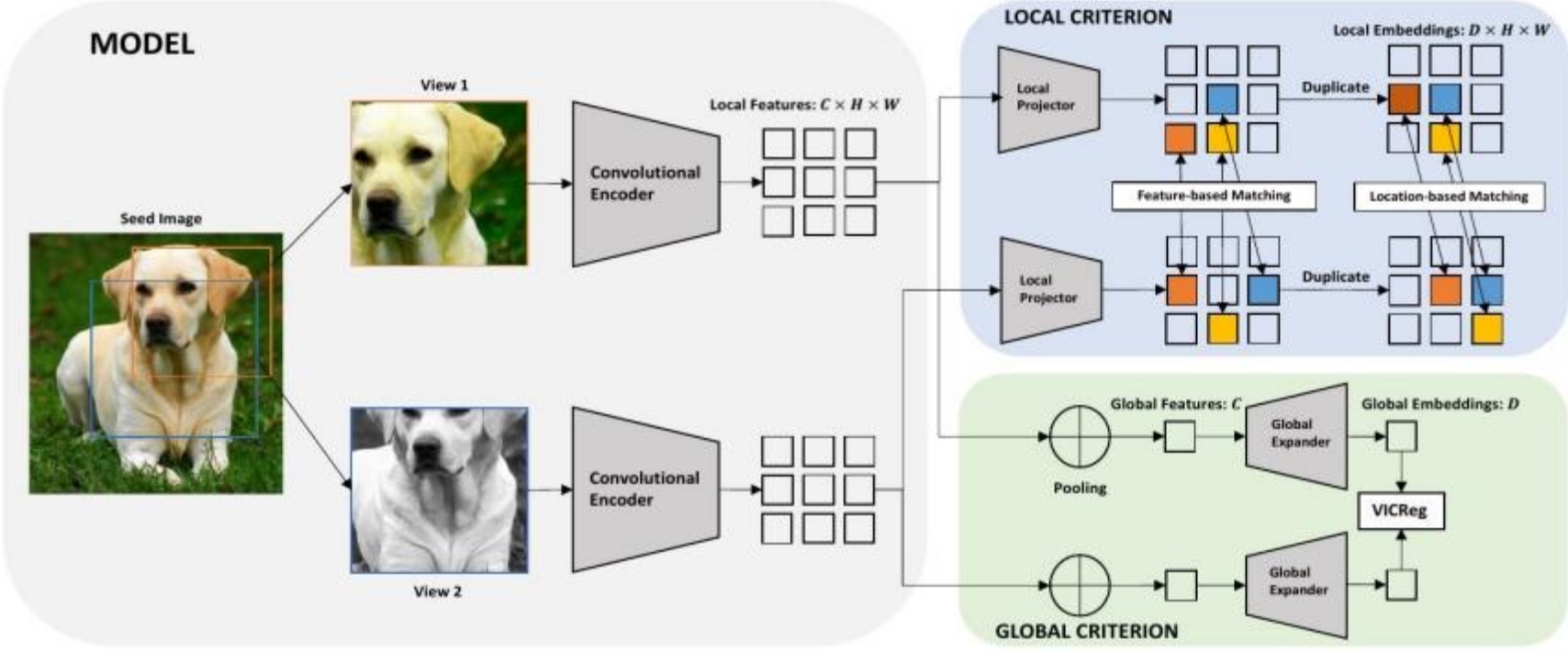
Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

VICReg：传输任务的结果。

Method	Linear Classification			Object Detection		
	Places205	VOC07	iNat18	VOC07+12	COCO det	COCO seg
Supervised	53.2	87.5	46.7	81.3	39.0	35.4
MoCo He et al. (2020)	46.9	79.8	31.5	-	-	-
PIRL Misra & Maaten (2020)	49.8	81.1	34.1	-	-	-
SimCLR Chen et al. (2020a)	52.5	85.5	37.2	-	-	-
MoCo v2 Chen et al. (2020c)	51.8	86.4	38.6	82.5	39.8	36.1
SimSiam Chen & He (2020)	-	-	-	82.4	-	-
BYOL Grill et al. (2020)	54.0	86.6	47.6	-	40.4†	37.0†
SwAV (m-c) Caron et al. (2020)	56.7	88.9	48.6	82.6	41.6	37.8
OBoW Gidaris et al. (2021)	56.8	89.3	-	82.9	-	-
Barlow Twins Grill et al. (2020)	54.1	86.2	46.5	82.6	40.0†	36.7†
VICReg (ours)	54.3	86.6	47.0	82.4	39.4	36.4

VICRegL：用于分割的局部匹配潜在变量

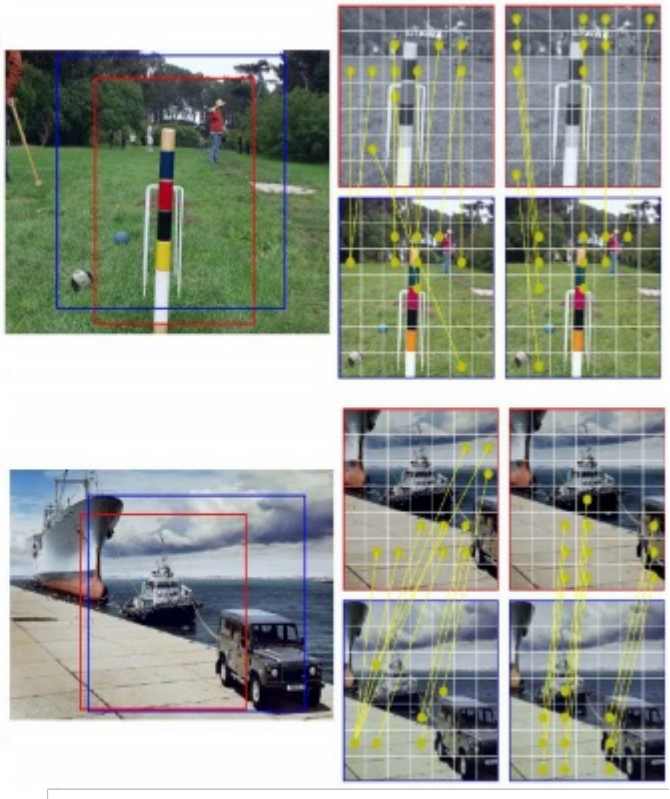
- **潜在变量优化：**
  - 查找两个图像的局部特征向量之间的配对
  - [Bardes, Ponce, LeCun, NeurIPS 2022, arXiv:2210.01571]



编译

VICRegL：用于分割的局部匹配潜在变量

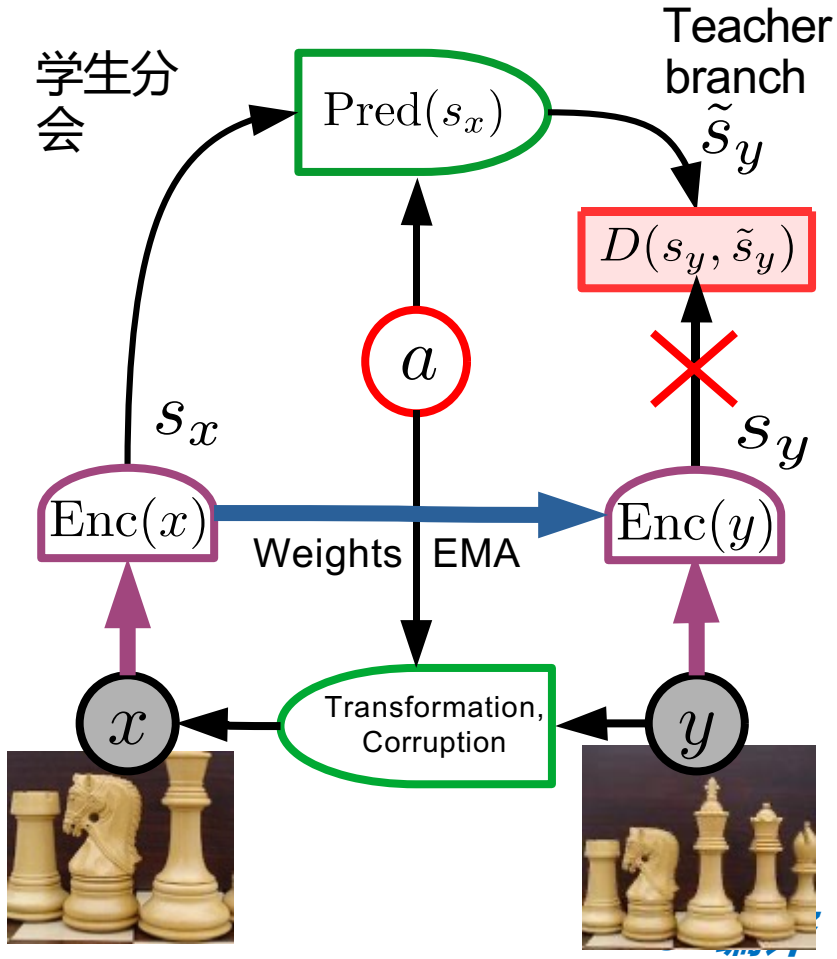
Method	Epochs	Linear Cls. (%) ImageNet Frozen	Linear Seg. (mIoU) Pascal VOC Frozen	Pascal VOC Fine-Tuned	Cityscapes Frozen
Global features					
MoCo v2 [Chen et al., 2020b]	200	67.5	35.6	64.8	14.3
SimCLR [Chen et al., 2020a]	400	68.2	45.9	65.4	17.9
BYOL [Grill et al., 2020]	300	**72.3**	47.1	65.7	22.6
VICReg [Bardes et al., 2022]	300	71.5	47.8	65.5	23.5
Local features					
PixPro [Xie et al., 2021]	400	60.6	52.8	67.5	22.6
DenseCL [Wang et al., 2021]	200	65.0	45.3	66.8	11.2
DetCon [Hénaff et al., 2021]	1000	66.3	53.6	67.4	16.2
InsLoc [Yang et al., 2022]	400	45.0	24.1	64.4	7.0
CP^2 [Wang et al., 2022]	820	53.1	21.7	65.2	8.4
ReSim [Xiao et al., 2021]	400	59.5	51.9	67.3	12.3
Ours					
VICRegL $\alpha = 0.9$	300	71.2	54.0	66.6	25.1
VICRegL $\alpha = 0.75$	300	70.4	**55.9**	**67.6**	**25.2**



CY编译

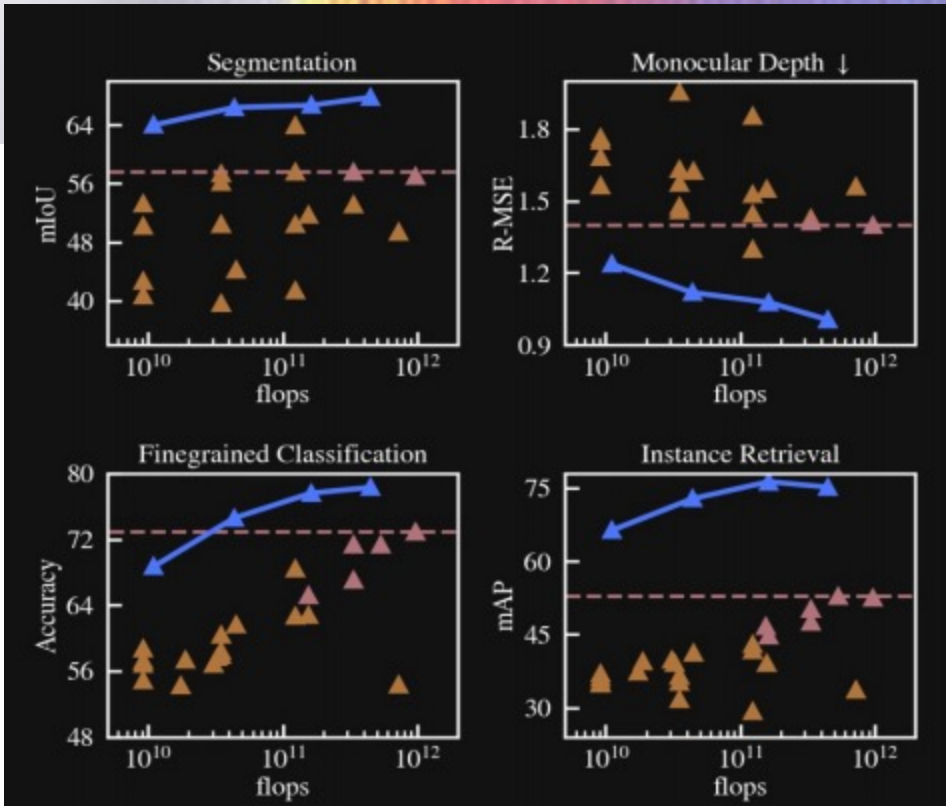
蒸馏方法

- **改良的暹罗网**
  - 预测器头消除了由于失真而导致的表示变化
- **例子:**
  - 引导你自己的潜伏 [Grill arXiv: 2006.07733]
  - SimSiam [Chen & He arXiv:2011.10566]
  - DINOv2 [Oquab arXiv:2304.07193]
- **优势**
  - 无阴性样本
- **缺点:**
  - 我们并不完全理解它为什么有效! [Tian et al. ArXiv:2102.06810]



DINOv2：图像基础模型

- **自监督通用图像特征**
- **Demo: https://dinov2.metademolab.com/**
- **Paper: [Oquab et al. ArXiv:2304.07193]**
  - 分类
    - 86.5% on IN1k with frozen features and linear head.
  - 细粒度分类
  - 深度估计
  - 语义分割
  - 实例检索
  - 密集和稀疏特征匹配



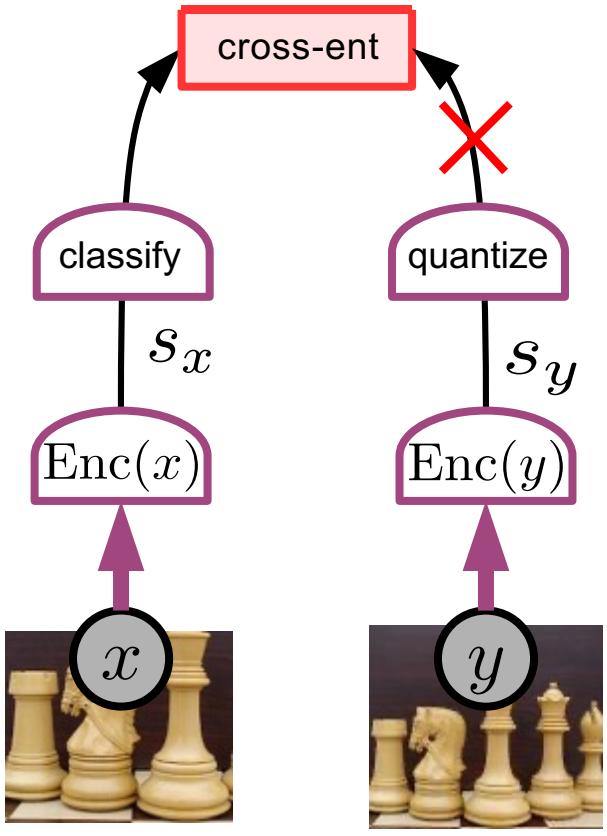
The DINOv2 family of models **drastically improves** over the previous state of the art in self-supervised learning (SSL), and **reaches performance comparable** with weakly-supervised features (WSL).

DINOv2：图像基础模型

- Demo: https://dinov2.metademolab.com/
- Paper: [Oquab et al. ArXiv:2304.07193]

DINOv2：联合嵌入架构

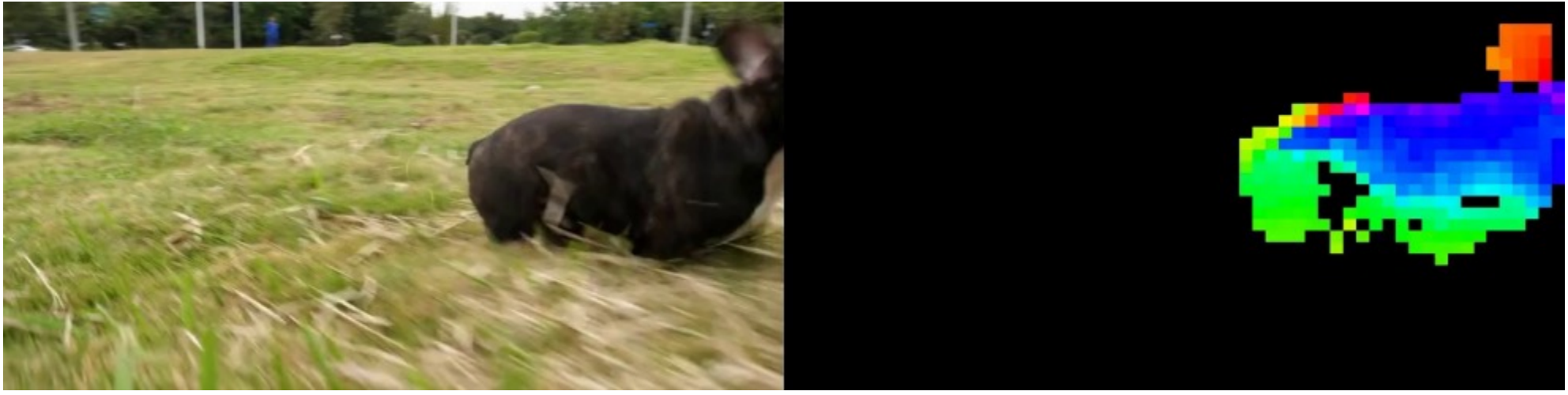
- 通过蒸馏进行SSL



Method	Arch.	Data	Text sup.	kNN val	linear val	linear ReaL	linear V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	×	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	×	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	×	–	79.8	–	–
MSN	ViT-L/7	INet-1k	×	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	×	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	×	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	×	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	×	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	×	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	×	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	×	**83.5**	**86.5**	**89.6**	**78.4**

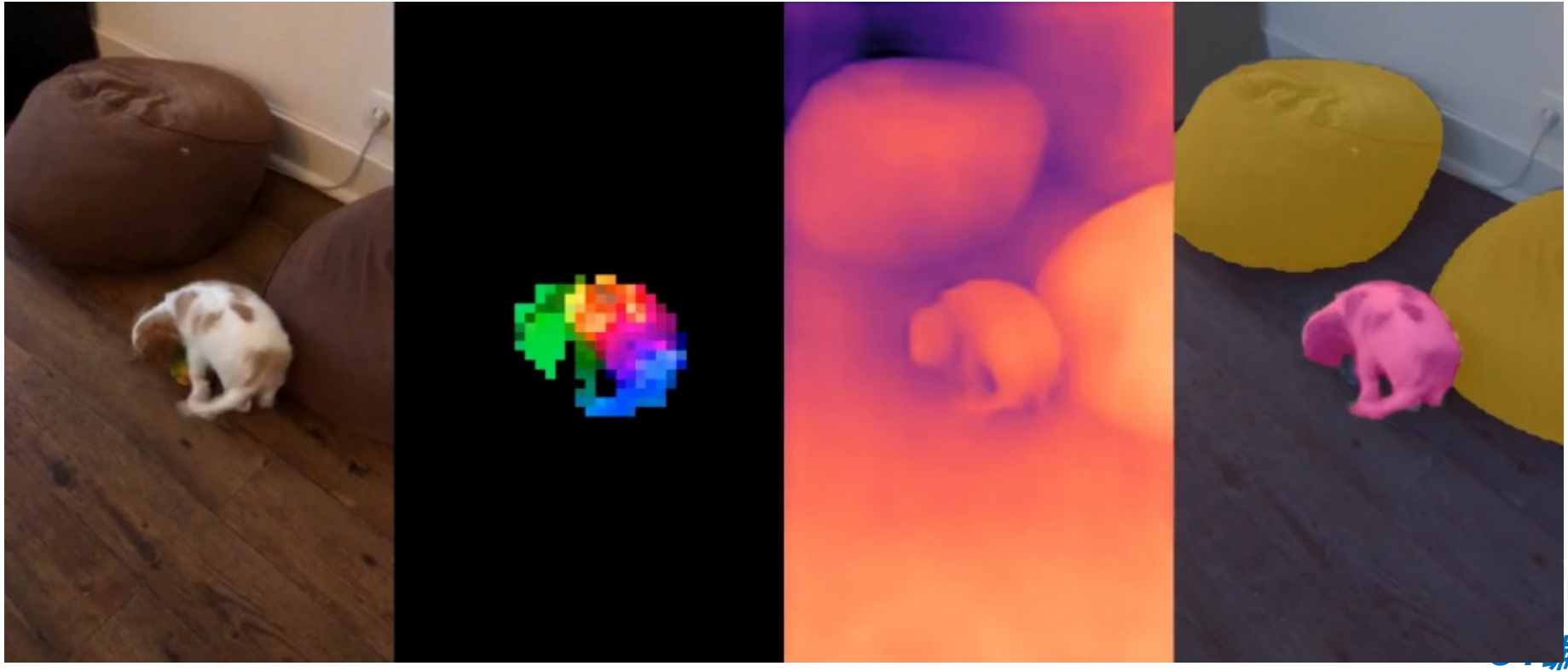
DINOv2

- 特征可视化：RGB = 前 3 个主成分



DINOv2

- 特征提取、深度估计、分割



使用 DINOv2 的树冠高度图

- **使用 DINOv2 特征从卫星图像估算树冠高度**
  - 使用激光雷达图像的地面实况
  - 0.5米分辨率图像
- **[ArXiv:2304.07213]**
  - Tolan 等人：亚米使用自监督学习和视觉转换器在空中和GEDI激光雷达上训练

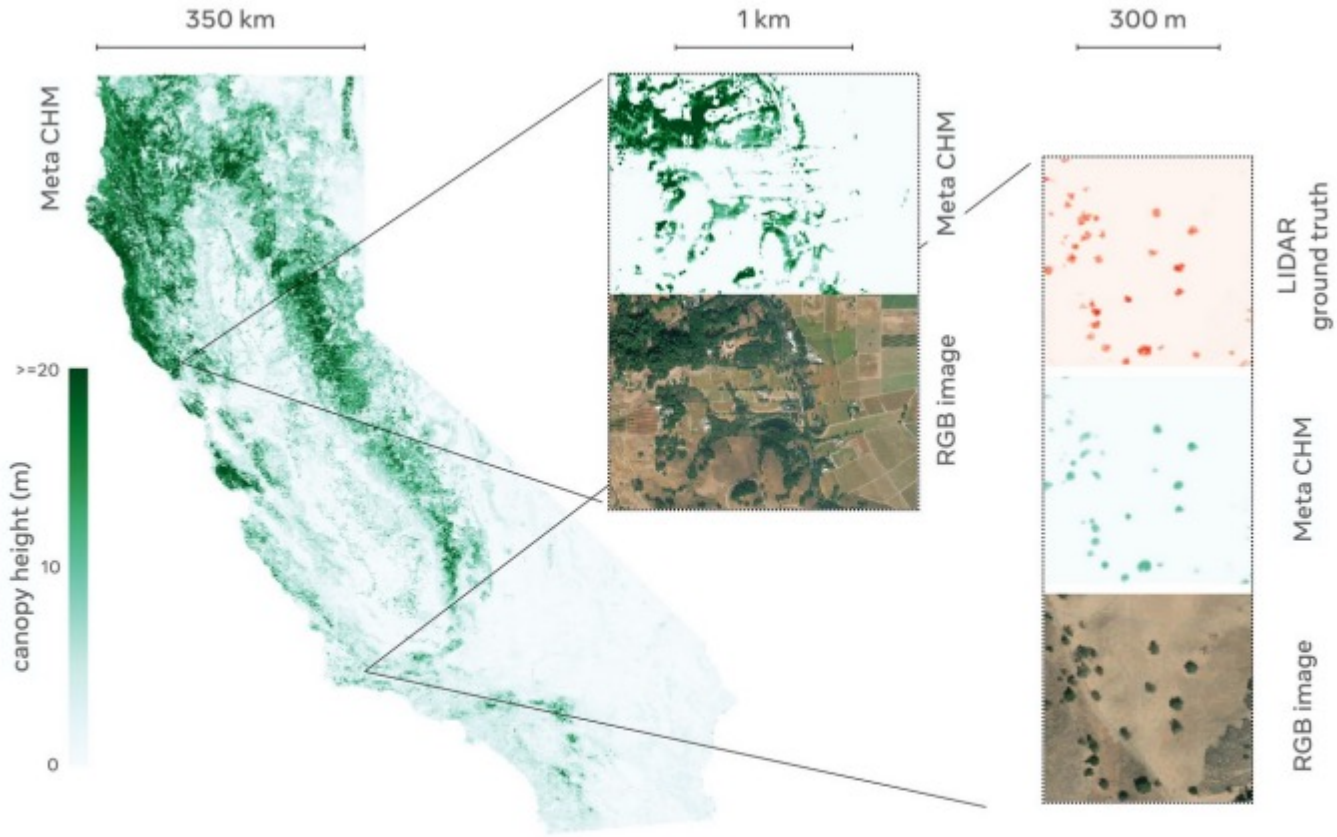
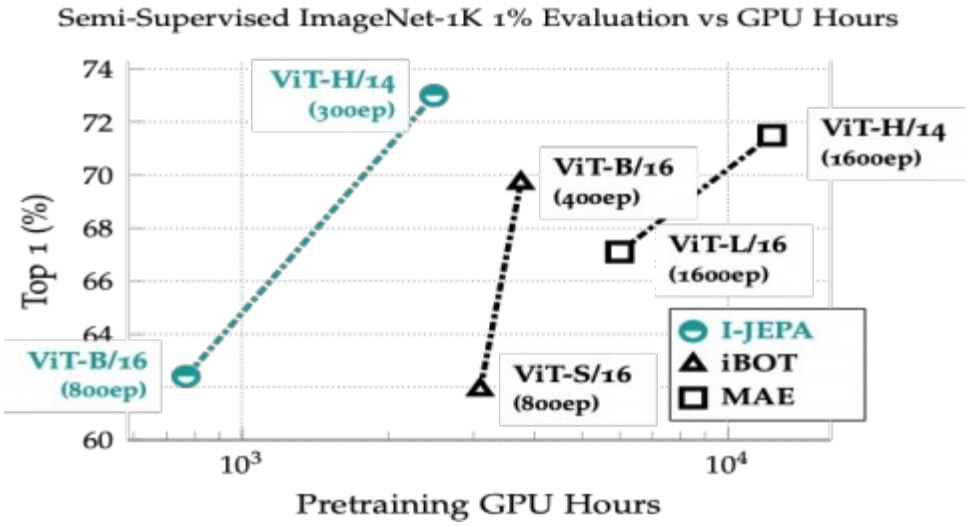
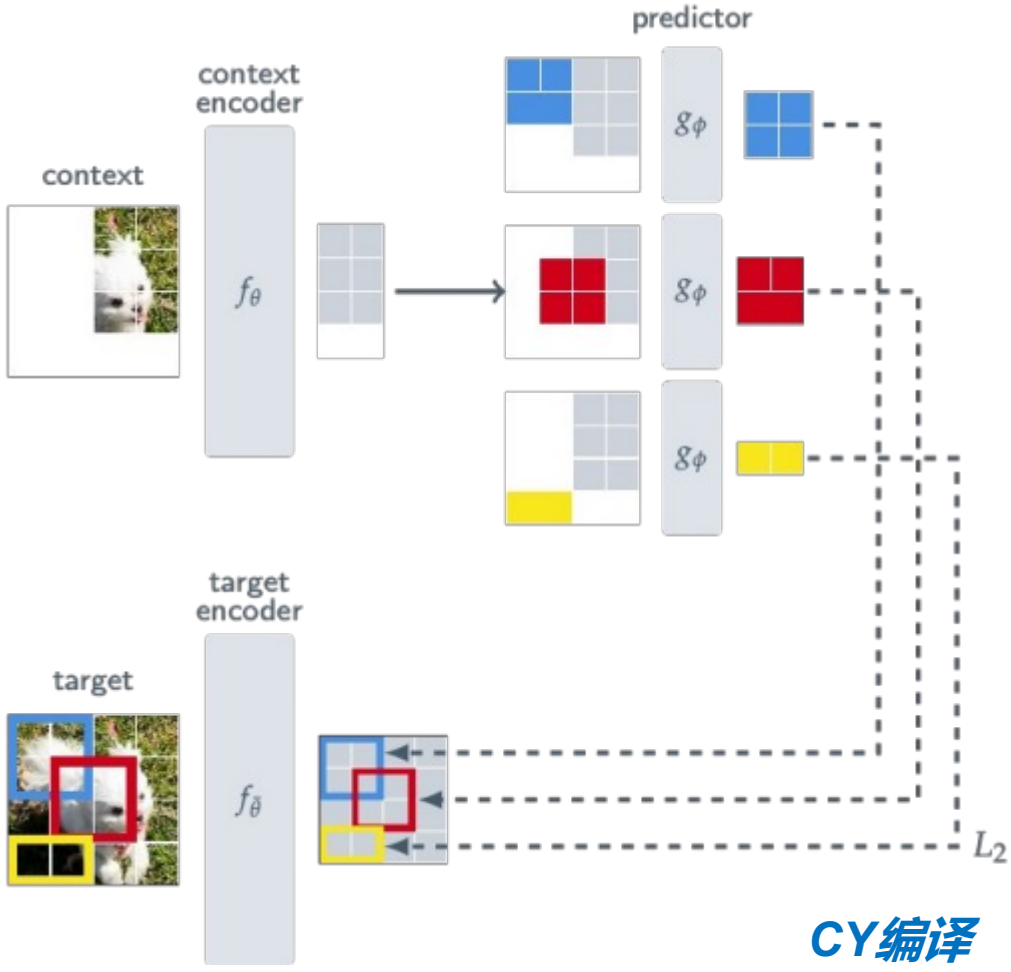


Figure 1: Canopy Height Map (CHM) for California, with inset showing zoomed-in region with input RGB imagery and LIDAR ground truth

Image-JEPA：使用屏蔽和转换器架构

- **“来自具有JEPA的图像的SSL”**
  - [M. Assran et al arxiv:2301.08243]
- **共同嵌入一个上下文和多个相邻补丁.**
  - 使用预测变量
  - 仅使用遮罩

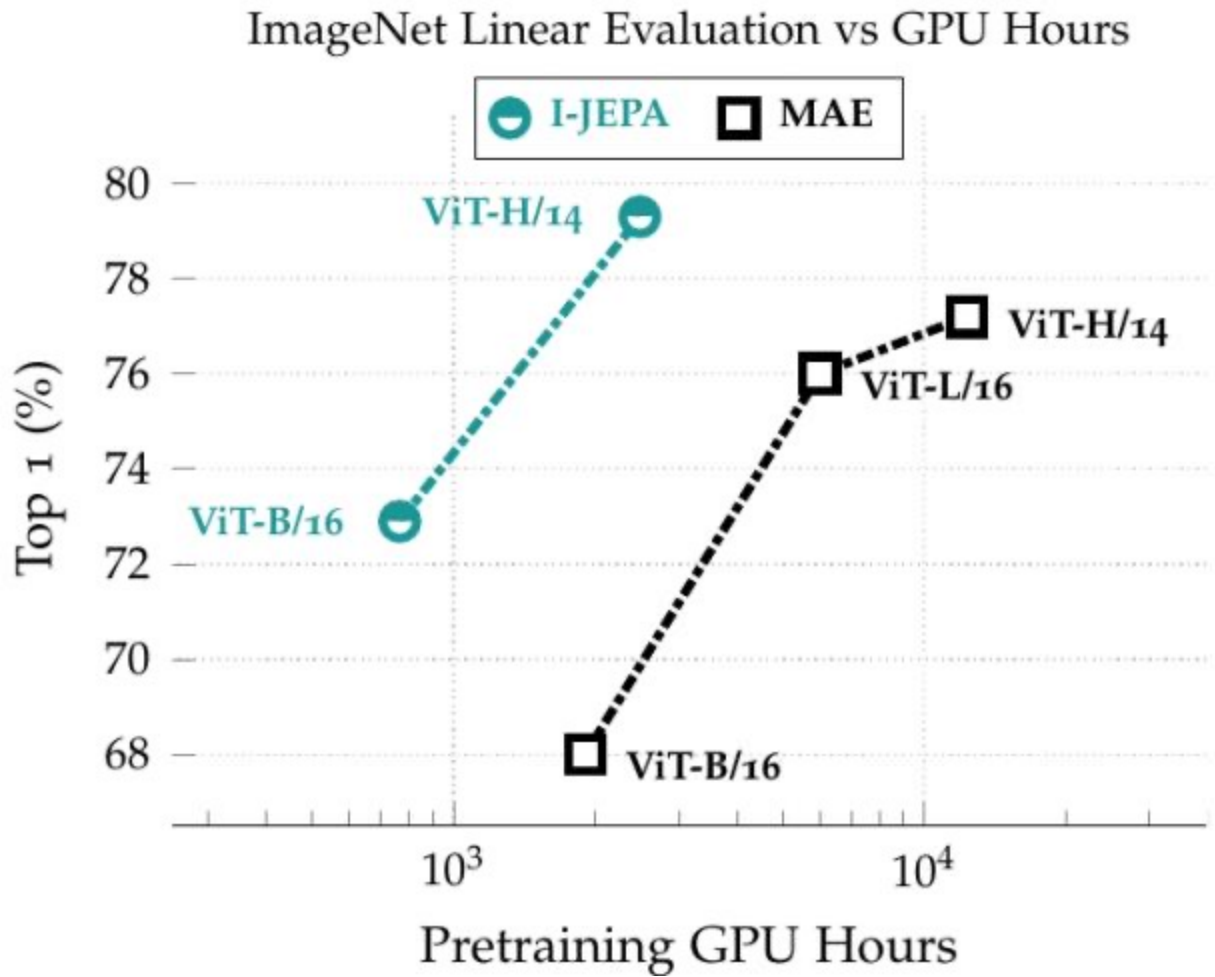




CY编译

I-JEPA结果

- 训练速度很快
- 非生成法节拍重建-基于生成方法，例如屏蔽自动编码器
  - (with a frozen trunk).



ImageNet上的I-JEPA结果

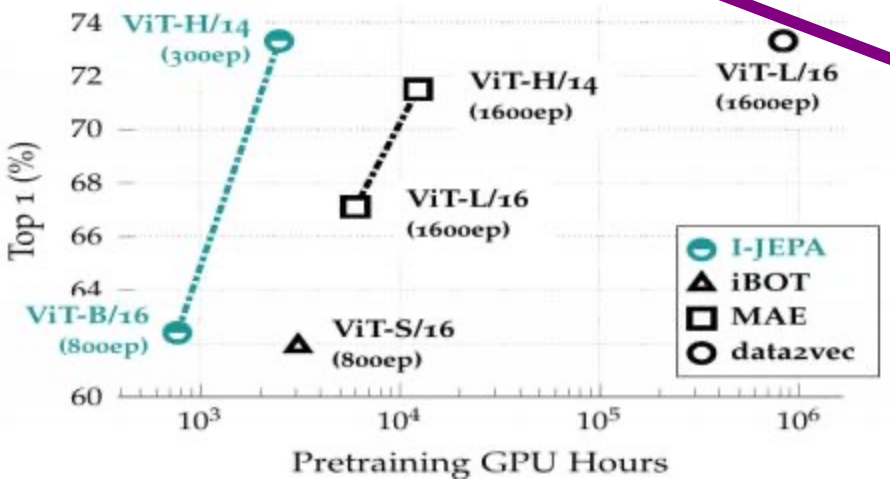
- JEPA优于生成式像素上的架构
- 使用数据增强的方法缩小差距
- 仅屏蔽的方法
  - 无数据增强
- 数据方法增大
  - 与SimCLR相似

Targets	Arch.	Epochs	Top-1
Target-Encoder Output	ViT-L/16	500	**66.9**
Pixels	ViT-L/16	800	40.7

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	53.5
MAE [34]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [20]	RN152 (2×)	800	79.1
DINO [17]	ViT-B/8	300	80.1
iBOT [74]	ViT-L/16	250	**81.0**

I-JEPA 在 ImageNet 上的结果，经过 1% 的训练

- JEPA在像素上优于生成式架构。
- 使用数据增强的方法缩小差距
- 仅屏蔽的方法
- 数据方法 增大



Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [7]	ViT-L/16	1600	73.3
MAE [34]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [74]	ViT-B/16	250	69.7
DINO [17]	ViT-B/8	300	70.0
SimCLR v2 [33]	RN151 (2×)	800	70.2
BYOL [33]	RN200 (2×)	800	71.2
MSN [3]	ViT-B/4	300	**75.7**

样本对比与维度对比?

- [Garrido et al. Arxiv：2206.02574， ICLR2023] （优秀论文，荣誉奖）
  “关于对比与非对比之间的二元性
- 对比自监督学习”

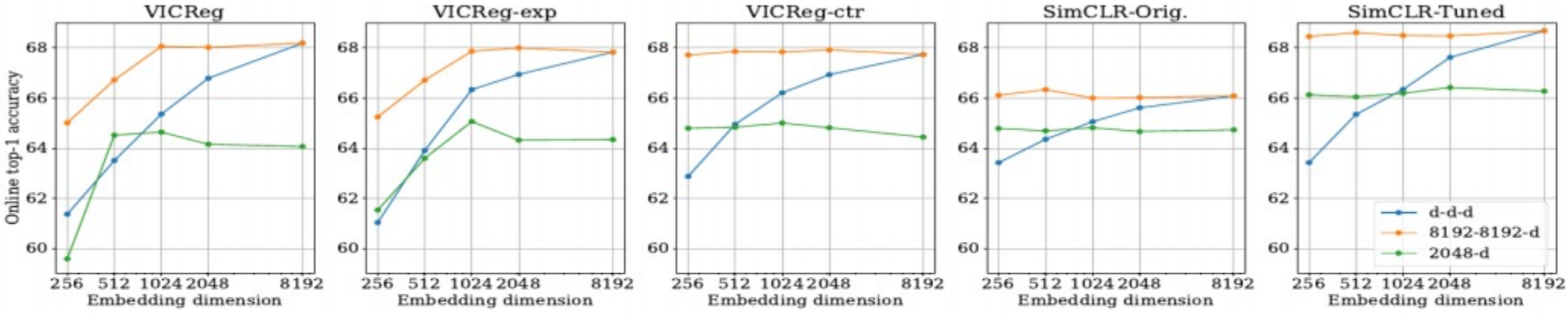


Figure 1: VICReg, VICReg-exp and VICReg-ctr perform similarly in 100 epochs training, validating empirically our theoretical result. While the original implementation of SimCLR performs significantly worse – which is unexpected per our theory – we are able to improve its performance to VICReg's level. This further validates our findings. While different projector architectures impact performance, behaviours are similar across methods. Confer supplementary section H for numerical values and hyperparameters.

Video-JEPA

https://github.com/facebookresearch/jepa

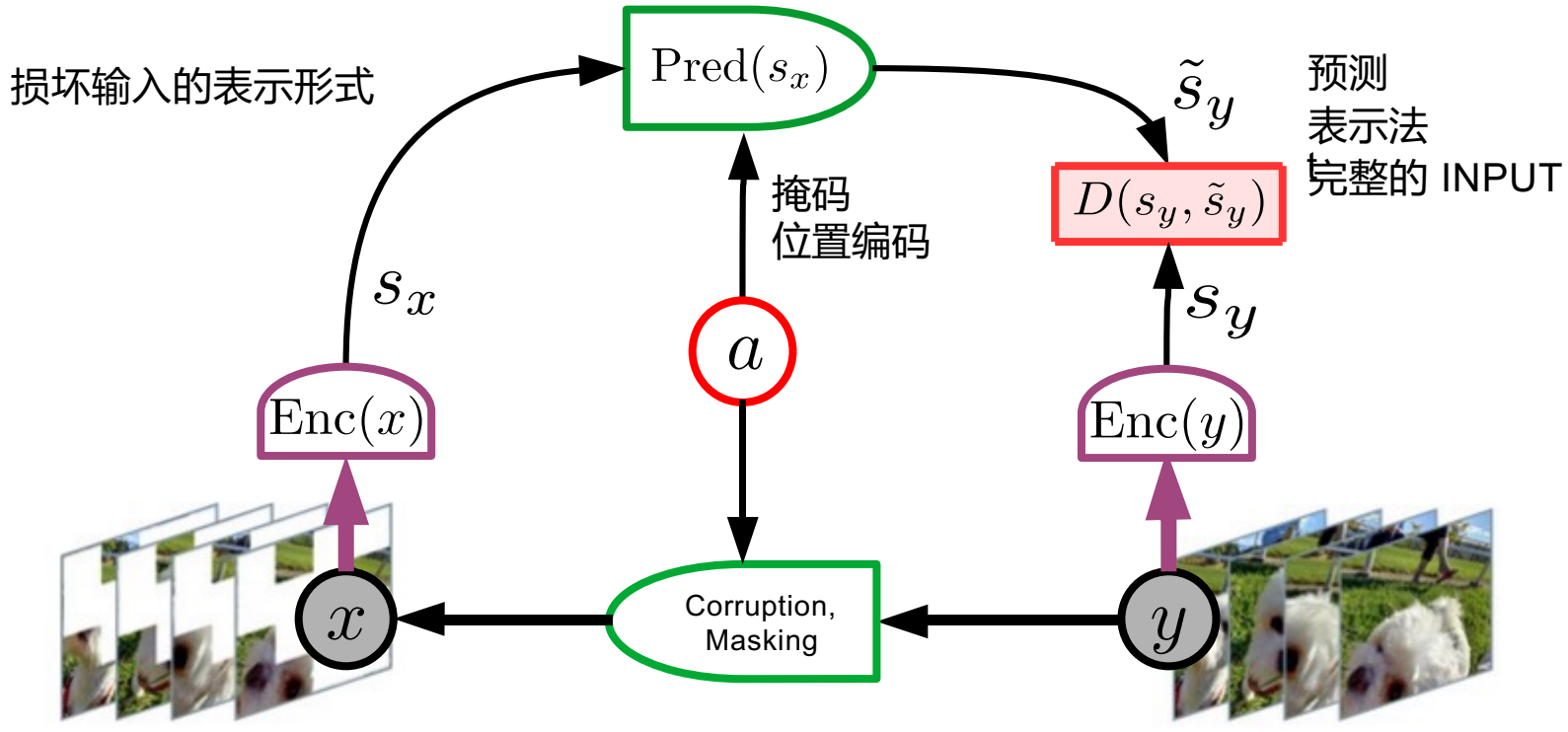
搜索 V-JEPA

“Revisiting Feature Prediction for Learning Visual Rerepresentations from Video” Adrien Bardes， Quentin Garrido， Jean Ponce， Xinlei Chen，Michael Rabbat，

扬·勒昆、马哈茂德·阿斯兰1、尼古拉斯·巴拉斯

Video-JEPA

- [Bardes et al. 2024]

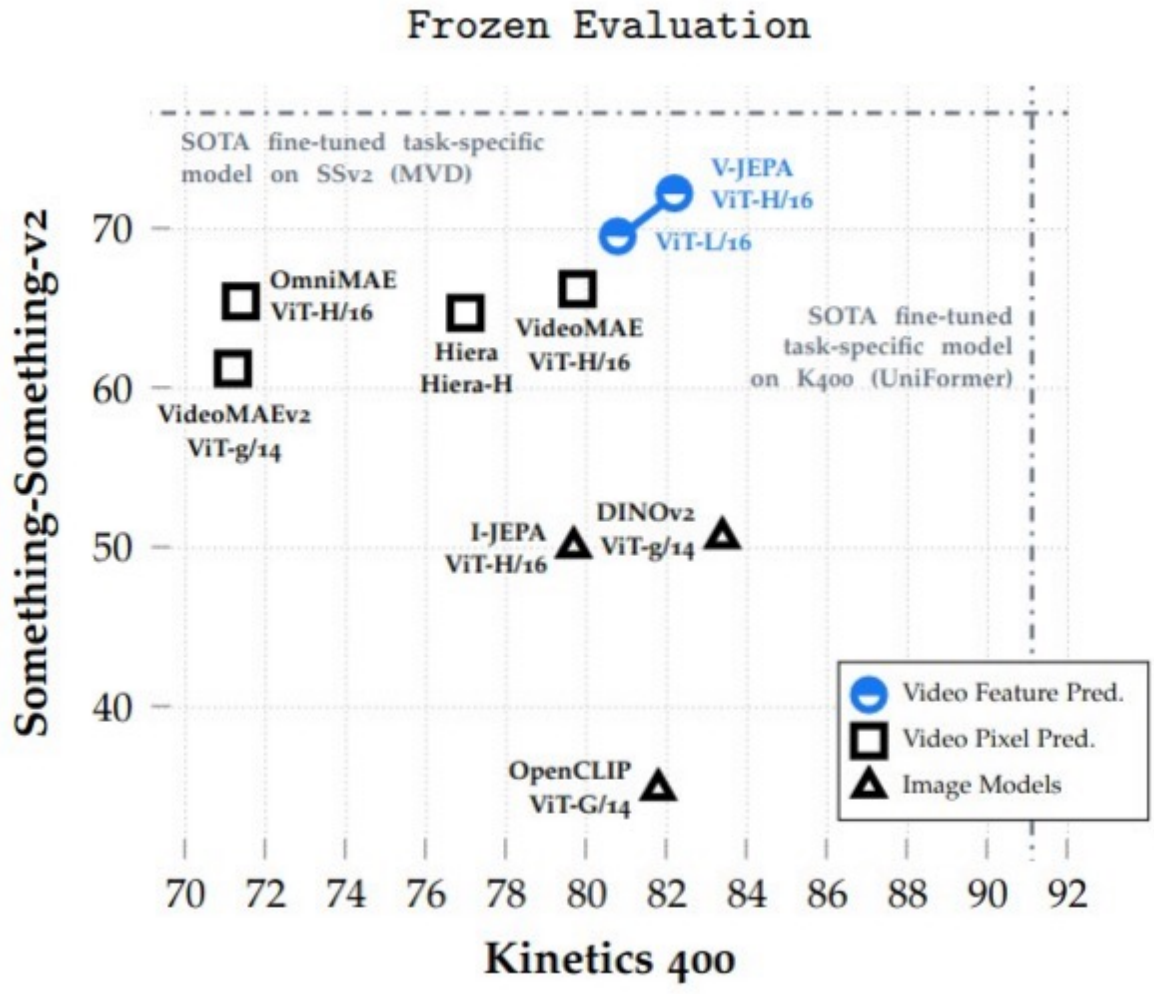


V-JEPA：动作识别结果

- 监督头部在冰冻的脊梁上。

  ]

- 比较
  生成模型：
  OmniMAE、VideoMAE、Hiera
- 与图像模型的比较：I-JEPA、DINOv2、OpenCLIP



V-JEPA:零样本动作识别的结果

- Rows 1-3:具有重构的衍生式架构
- Row 4: V-JEPA
- 监督头部在冰冻的脊梁上。

		Frozen Evaluation					
		K400 (16×8×3)			SSv2 (16×2×3)		
Method	Arch.	5%	10%	50%	5%	10%	50%
MVD	ViT-L/16	62.6 ± 0.2	68.3 ± 0.2	77.2 ± 0.3	42.9 ± 0.8	49.5 ± 0.6	61.0 ± 0.2
VideoMAE	ViT-H/16	62.3 ± 0.3	68.5 ± 0.2	78.2 ± 0.1	41.4 ± 0.8	48.1 ± 0.2	60.5 ± 0.4
VideoMAEv2	ViT-g/14	37.0 ± 0.3	48.8 ± 0.4	67.8 ± 0.1	28.0 ± 1.0	37.3 ± 0.3	54.0 ± 0.3
V-JEPA	ViT-H/16_{384}	68.2 ± 0.2	72.8 ± 0.2	80.6 ± 0.2	54.0 ± 0.2	59.3 ± 0.5	67.9 ± 0.2

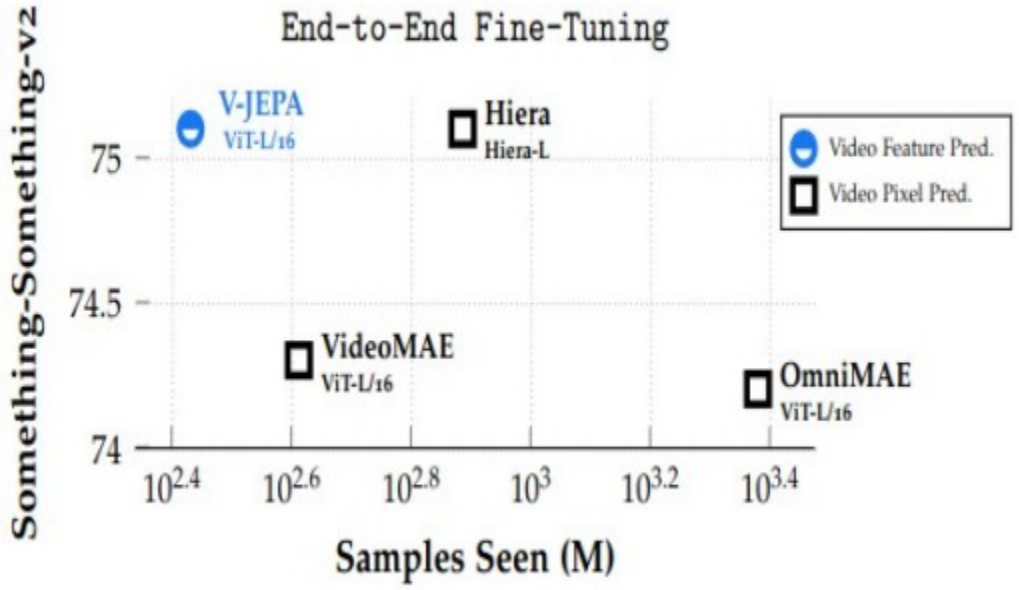
V-JEPA:视频训练与图像训练

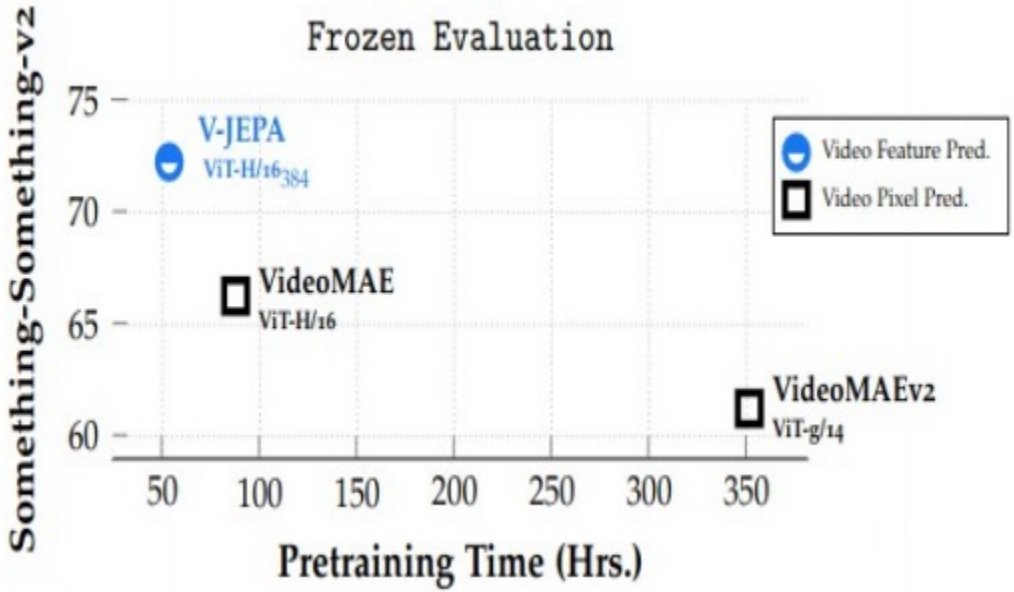
- 冷冻评估
- 视频预训练在动作识别方面效果更好
- V-JEPA:在视频模型中，ImageNet1K的最佳效果

Method	Arch.	Params.	Data	*Video Tasks* K400 (16×8×3)	SSv2 (16×2×3)	AVA	*Image Tasks* IN1K	Places205	iNat21
Methods pretrained on Images									
I-JEPA	ViT-H/16_{512}	630M	IN22K	79.7	50.0	19.8	84.4	66.5	85.7
OpenCLIP	ViT-G/14	1800M	LAION	81.8	34.8	23.2	85.3	**70.2**	83.6
DINOv2	ViT-g/14	1100M	LVD-142M	**83.4**	50.6	24.3	**86.2**	68.4	**88.8**
Methods pretrained on Videos									
MVD	ViT-L/16	200M	IN1K+K400	79.4	66.5	19.7	73.3	59.4	65.7
OmniMAE	ViT-H/16	630M	IN1K+SSv2	71.4	65.4	16.0	76.3	60.6	72.4
VideoMAE	ViT-H/16	630M	K400	79.8	66.2	20.7	72.3	59.1	65.5
VideoMAEv2	ViT-g/14	1100M	Un.Hybrid	71.2	61.2	12.9	71.4	60.6	68.3
Hiera	Hiera-H	670M	K400	77.0	64.7	17.5	71.4	59.5	61.7
V-JEPA	ViT-L/16	200M	VideoMix2M	80.8	69.5	25.6	74.8	60.3	67.8
	ViT-H/16	630M		**82.0**	71.4	**25.8**	75.9	61.7	67.9
	ViT-H/16_{384}	630M		81.9	**72.2**	25.0	**77.4**	**62.8**	**72.6**

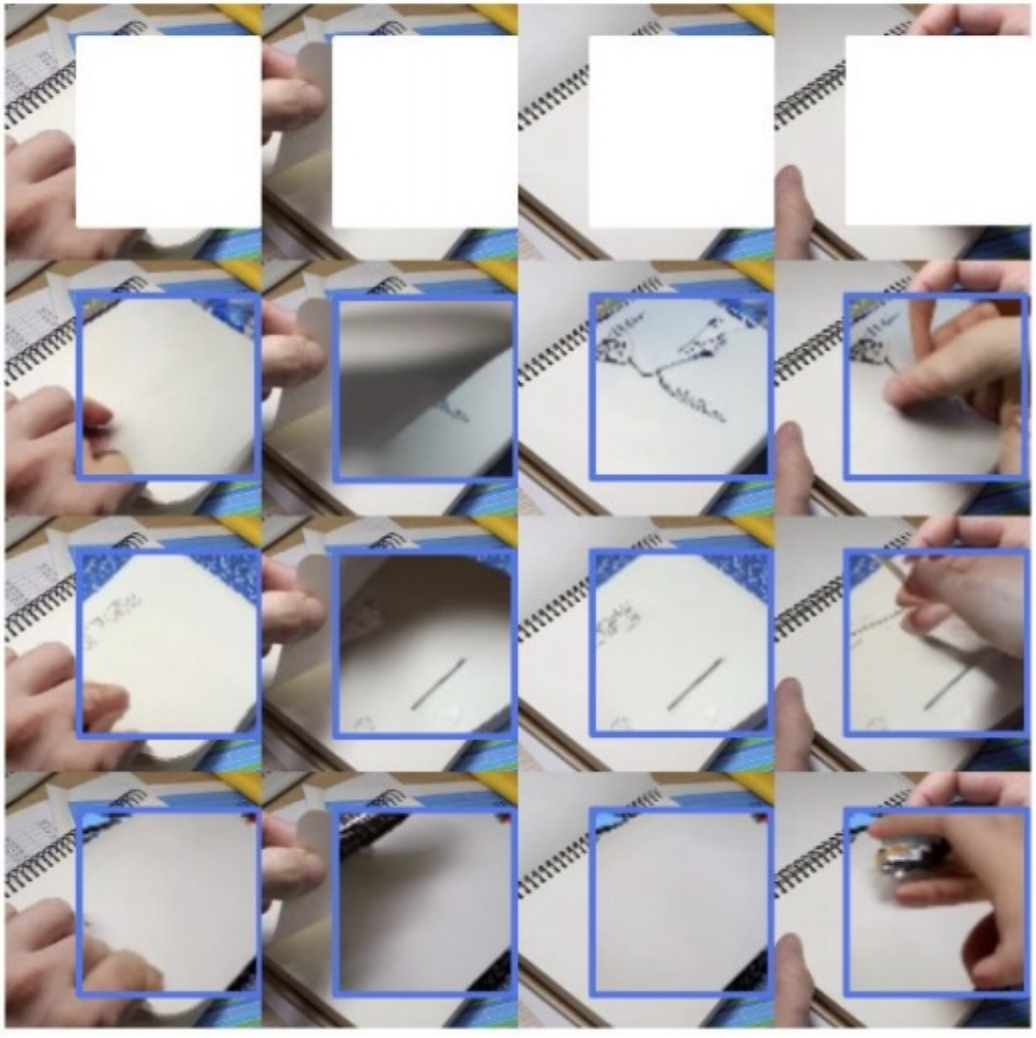
V-JEPA:采样效率和学习速度

- 对 Something-Something-v2 的评估
- 与基于重建的生成方法的比较





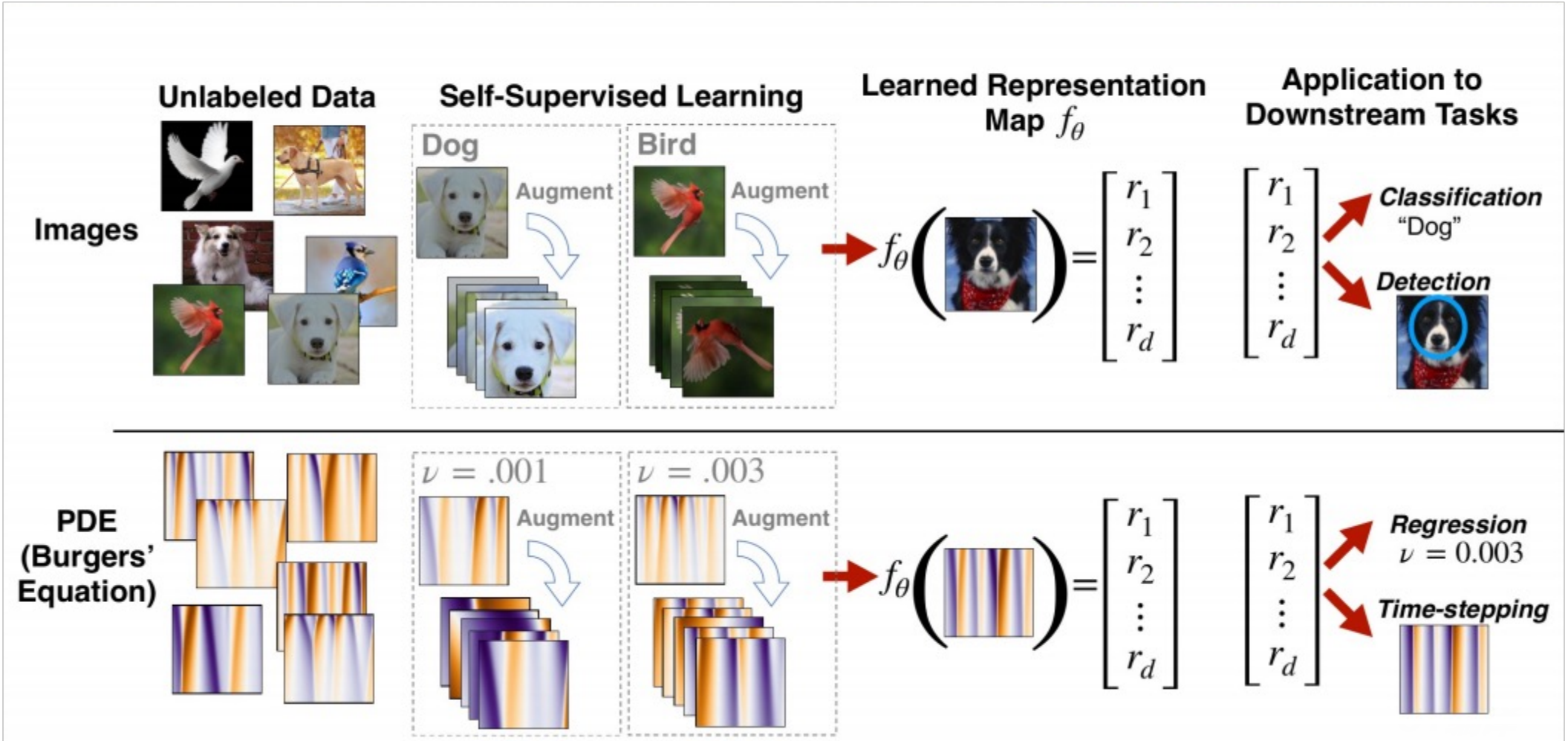
V-JEPA:使用单独训练的解码器进行重建



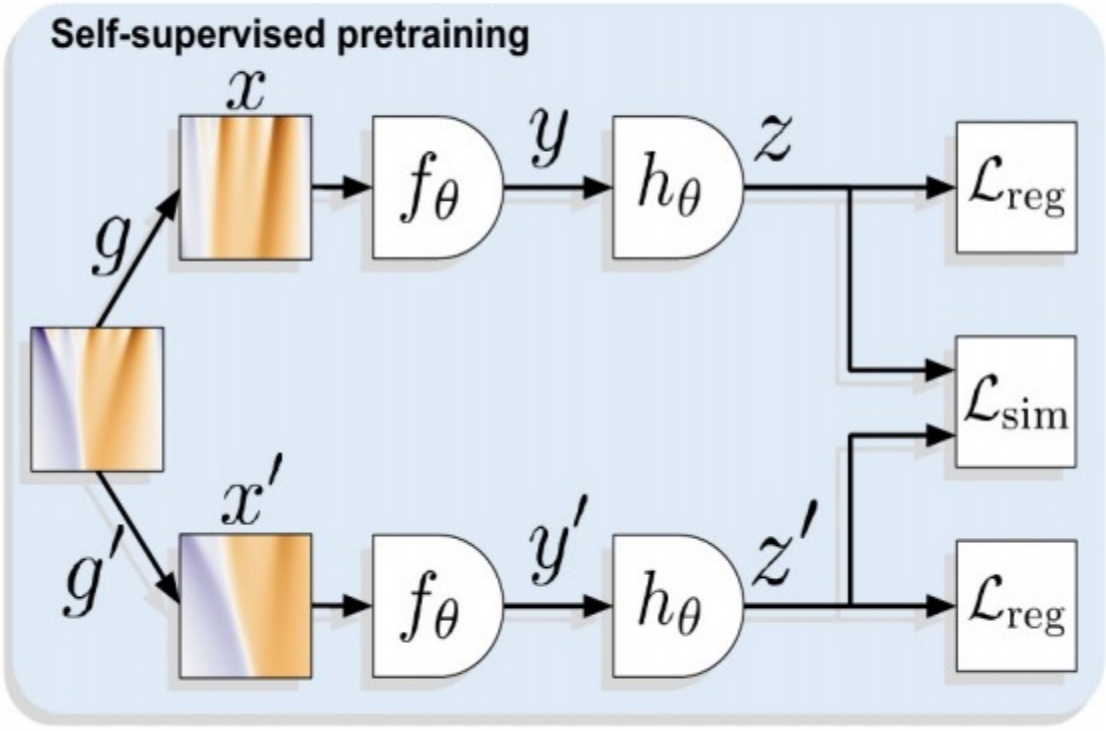
SSL for PDEs

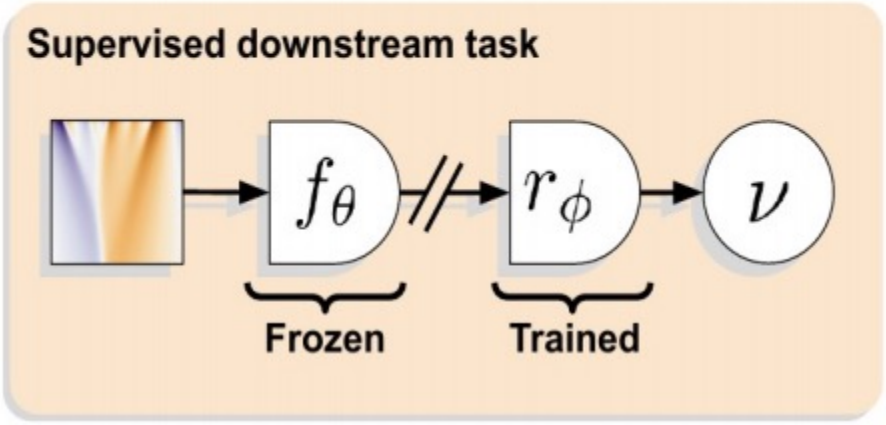
ArXiv:2307.05432　NeurIPS 2023
偏微分方程的李对称自监督学习 Grégoire Mialon， Quentin Garrido， Hannah Lawrence， Danyal Rehman， Yann LeCun， Bobak T. Kiani

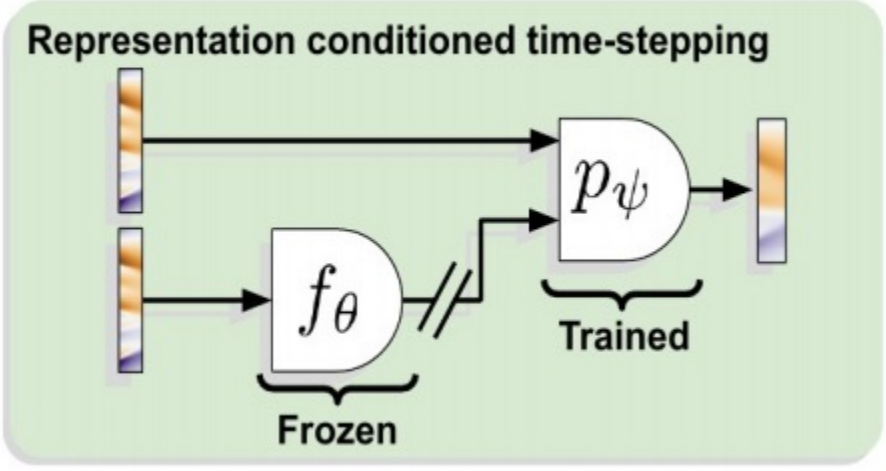
SSL for PDE:使用 VICReg 提取动态参数



使用 VICReg 学习方程的表示。







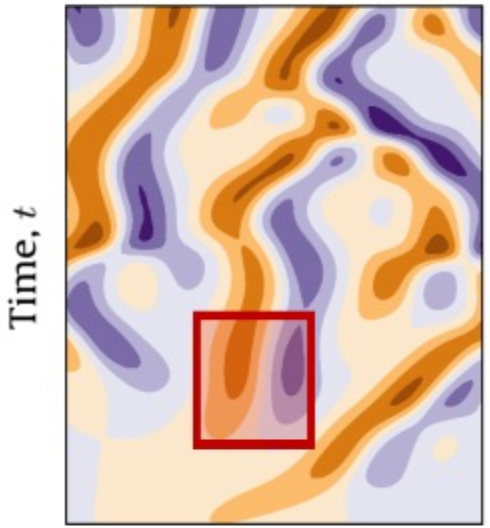
SSL for PDE

An example: the **Kuramoto-Sivashinsky (KS)** equation is a model of chaotic flow given by

$$u_t + uu_x + u_{xx} + u_{xxxx} = 0,$$

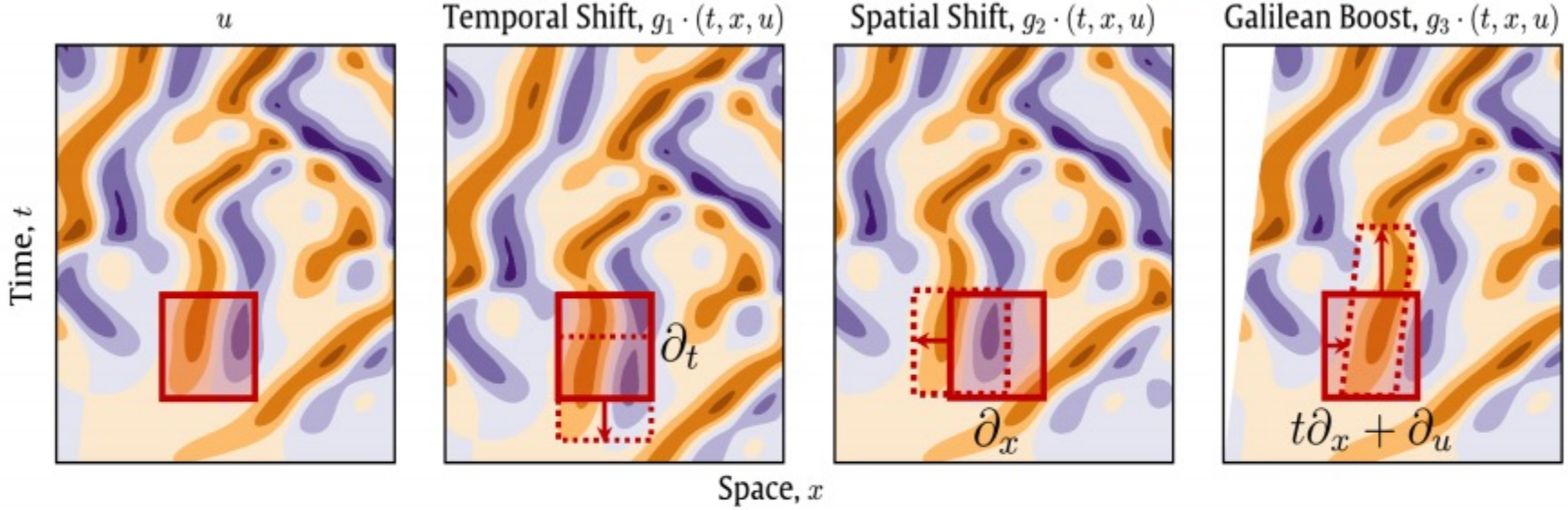
where $u(x, t)$ is the dependent variable.

- Often shows up in reaction-diffusion systems or flame propagation problems.
- Solution can be seen as an image...
- Admit Lie point symmetries: smooth transformations of a solution producing another solution to the same PDE.
- Can be used to learn models [Brandstetter et al., 2022].



A 1D solution to KS (x-axis is space).

SSL for PDE: Data "augmentation"



One parameter Lie point symmetries for the Kuramoto-Sivashinsky (KS) PDE. Left to right: un-modified solution (u), temporal shifts (g_1), spatial shifts (g_2), and Galilean boosts (g_3) with corresponding infinitesimal transformations in the Lie algebra placed inside the figure. The shaded red square denotes the original (x, t), while the dotted line represents the same points after the augmentation is applied.

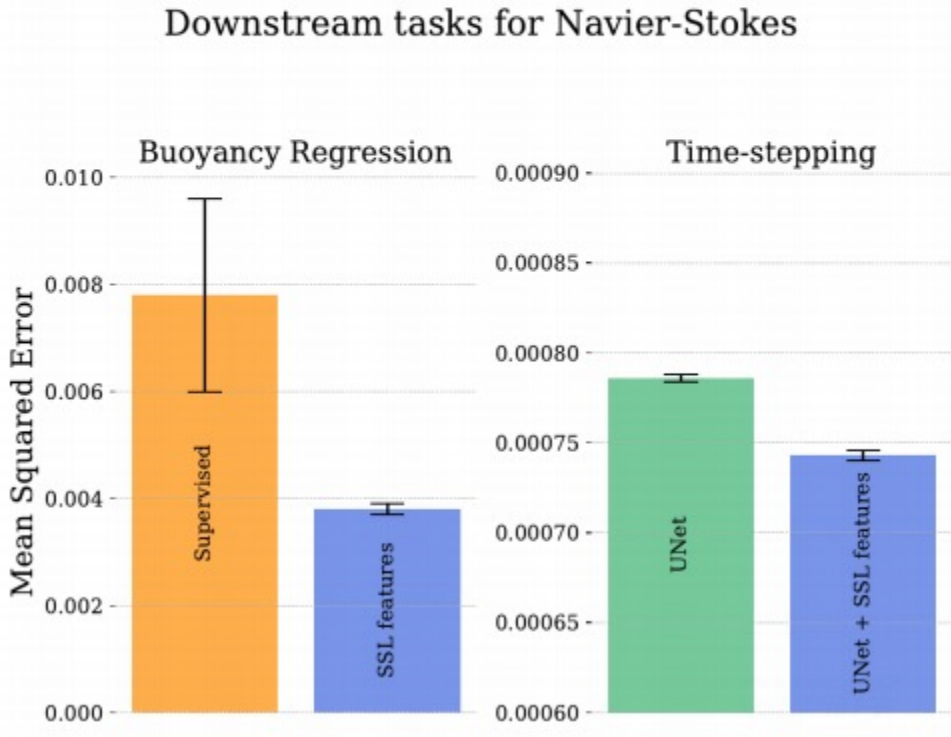
$$\begin{aligned} \text{Temporal Shift:} \quad & g_1(\epsilon) : (x, t, u) \mapsto (x, t + \epsilon, u) \\ \text{Spatial Shift:} \quad & g_2(\epsilon) : (x, t, u) \mapsto (x + \epsilon, t, u) \\ \text{Galilean Boost:} \quad & g_3(\epsilon) : (x, t, u) \mapsto (x + \epsilon t, t, u + \epsilon) \end{aligned}$$

SSL用于预测纳维-斯托克斯的浮力

The **incompressible Navier-Stokes** equation is given by

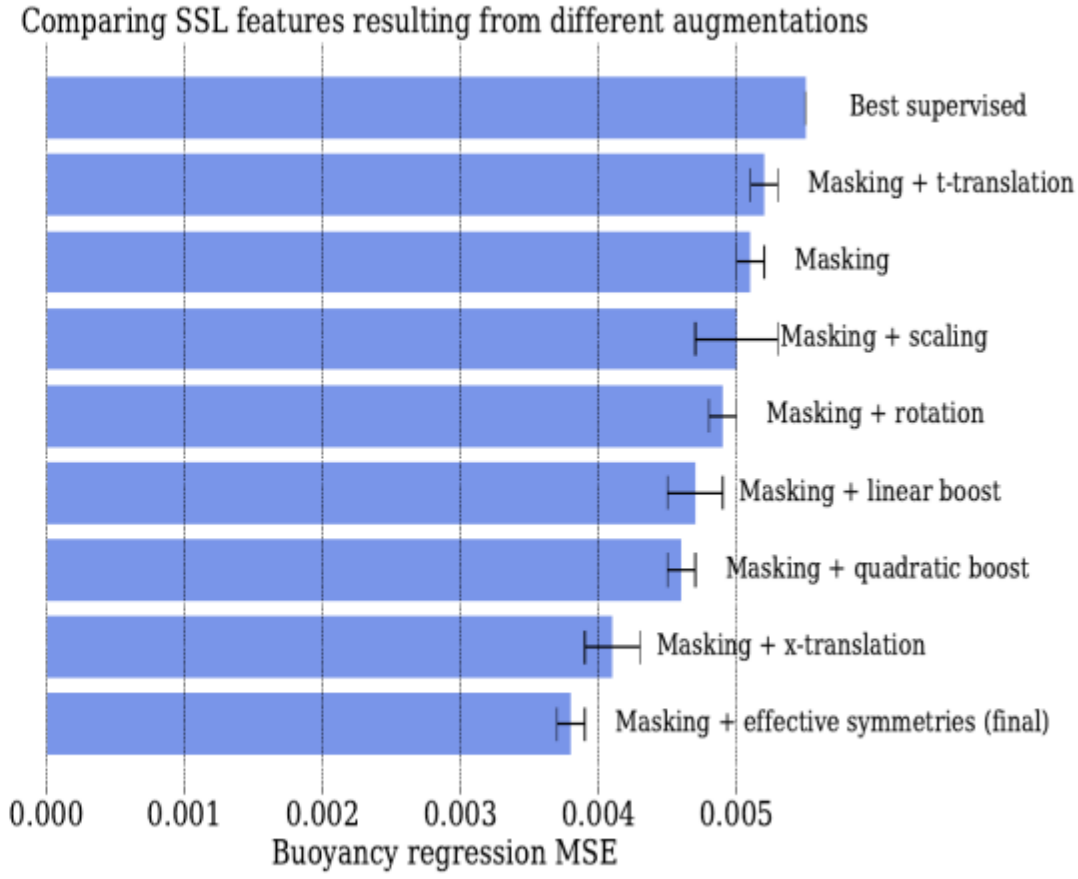
$$\boldsymbol{u}_t = -\boldsymbol{u} \cdot \nabla \boldsymbol{u} - \frac{1}{\rho} \nabla p + \nu \nabla^2 \boldsymbol{u} + \boldsymbol{f}, \quad \nabla \boldsymbol{u} = 0.$$

- 26k 2D trajectories, 56 frames (128x128) each [Gupta and Brandstetter, 2023].
- Task 1: regressing buoyancy $\boldsymbol{f}$.
- Task 2: Time-stepping, predict next frames given past frames.
- SSL features are effective and easy to use.



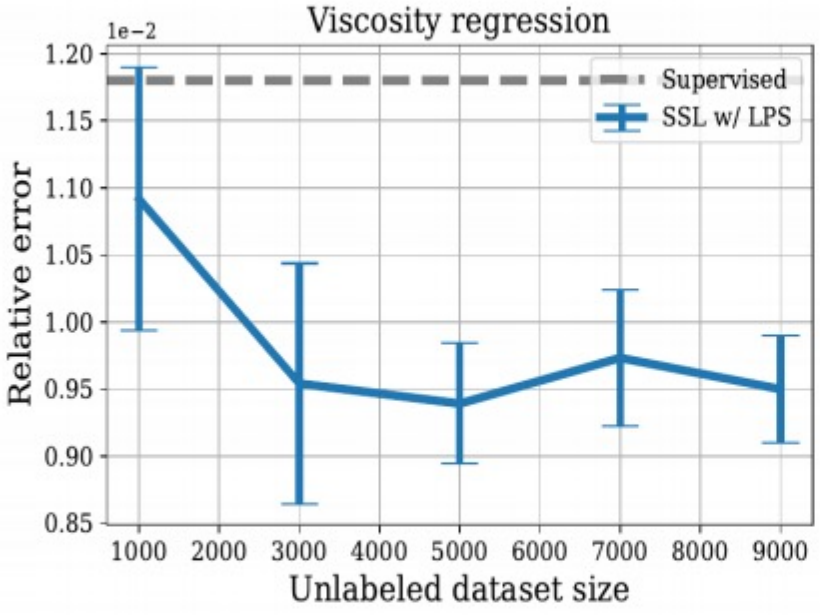
SSL用于预测纳维-斯托克斯的浮力

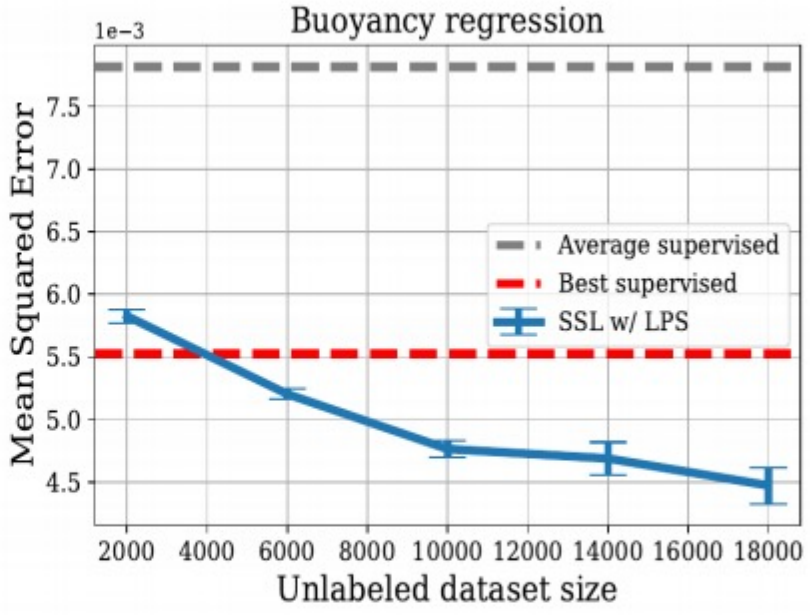
- Navier-Stokes: 8 Lie symmetrie groups, with varying strength.
- Intuition is not sufficient to select augmentations.
- Optimal mix is different from supervised [Brandstetter et al., 2022].
- Masking is necessary but not really sufficient.



SSL pre-training gives better results than purely supervised

SSL vs. supervised: open question in vision [Sariyildiz et al., 2023, Oquab et al., 2023]. Here, big discrepancy.





Influence of dataset size on regression tasks. **(Left)** Kinematic regression on Burger's equation. **(Right)** Buoyancy regression on Navier-Stokes' equation.

要解决的问题

- **基于能量的学习的数学基础**
  - 能量表面的几何形状、缩放定律、边界...
- **具有正则化潜在变量的 JEPA**
  - 在非确定性环境中的学习和规划
- **存在不确定性的规划算法**
  - 基于梯度的方法和组合搜索方法
- **学习成本模块（反向 RL）**
  - 基于能量的方法：为观测到的轨迹提供低成本
- **使用不准确的世界模型进行规划**
  - 防止在空间的不确定部分制定不良计划
- **探索调整世界模型**
  - 好奇心的内在目标

我们正在做的事情

- **视频自监督学习**
  - 分层视频-使用SSL训练的JEPA
- **能够推理和计划的LLM，由目标驱动**
  - 在表示空间中规划并使用 AR-LLM 将表示转换为文本的对话系统
- **学习分层规划**
  - 就玩具规划问题对多时间尺度的 H-JEPA 进行训练。

点

- **计算能力**
  - AR-LLM 对每个令牌使用固定数量的计算量
  - 目标驱动的 AI 是图灵完备的（推理 == 优化）
- **我们仍然缺少达到人类水平人工智能的基本概念**
  - 扩大自回归 LLM 不会把我们带到那里
  - 我们需要机器来了解世界是如何运作的
- **具有自监督学习和 JEPA 的学习世界模型**
  - 非生成架构，在表示空间中预测
- **目标驱动的 AI 架构**
  - 可以计划他们的答案
  - 必须满足目标：可操纵和可控
  - 护栏物镜可以通过施工确保安全。

未来的通用虚拟助手

- **我们与数字世界的所有互动将由人工智能助手调解。**
  - 它们将构成所有内容的存储库人类知识与文化
  - 它们将构成一个共享的基础设施就像今天的互联网一样。

- **这些 AI 平台必须是开源的**
  - 否则，我们的文化将被美国西海岸或中国的少数公司控制。
  - 训练他们必须通过众包方式进行
- **开源 AI 平台是必要的**

这一愿景对政策意味着什么？

- **人工智能系统将成为通用平台**
- **平台（基础模型）将是开源的**
  - 它们将凝聚人类所有的知识
  - 为了安全起见，将共享护栏目标
- **训练和微调将采用众包方式**
  - 语言、文化和利益集团将对基本模型进行微调，以满足他们的兴趣。

- **垂直应用的专有系统将建立在顶部**
- **当每个人都有一个人工智能助手时，我们将需要**
  - 用于推理的海量计算基础设施：高效推理芯片.
- **开源人工智能绝不能被监管**
  - **AI 联盟：Meta、NAB、索尼、索尼、国家、开端.....**

问题

- **达到人类水平的人工智能需要多长时间?**
  - 数年到数十年。途中有许多问题需要解决。
  - 在进入HLAI之前，我们将进入猫级AI，狗级AI,...
- **什么是AGI?**
  - 没有这样的事情。智能是高度多维的
  - 智力是技能+快速学习新技能的能力的集合
  - 即使是人类也只能完成所有任务的一小部分
- 机器会超越人类智能吗?
  - 是的，他们已经在一些狭窄的领域这样做了。
  - **毫无疑问，机器最终将在人类智能（甚至更多）的所有领域超越人类智能**

问题

- **强大的人工智能是否存在短期风险?**
  - 是的，就像每一项技术一样。
  - 虚假信息、宣传、仇恨、垃圾邮件,...：人工智能就是解决方案!
  - 信息来源的集中
  - 所有这些风险都可以减轻
- **（超）人类水平的人工智能是否存在长期风险?**
  - 机器人不会接管世界！人性在机器上的错误投射
  - 智力与支配欲望无关，即使在人类中也是如此
  - 目标驱动的人工智能系统将屈从于人类
  - 人工智能不会是与我们竞争的“物种”。
  - 我们将设计其目标和护栏。

Questions

- 如何解决对齐问题?
  - 通过在沙盒系统中进行反复试验和测试
  - 我们非常熟悉为人类和超人实体设计目标。这就是所谓的立法。
    如果坏人掌握了强大的人工智能怎么办?
  - 他们的邪恶 AI 将被好人的 AI 警察击倒。
- 人类水平的人工智能有什么好处?

- 人工智能将放大人类智能，进步将加速
- 就好像每个人都有一个超级　聪明的员工为他们工作

- 对社会的影响可能与印刷机一样深远
- 通过放大人类智能，人工智能将带来启蒙的新时代，人类的新复兴.

Thank you!

NEW YORK UNIVERSITY

Meta AI

个人自媒体：苏哲管理咨询

CY编译