

SimCLR Framework

✓ 想想这一年来你最常听到过哪些词

✎ 自监督学习，对比学习等，这些事好像都不需要我们准备标签

✎ Openai开创了GPT系列，CLIP，Dalle等，都在告诉我们一件事

✎ 模型在训练的时候，不要被标签所束缚，模型的潜力应该由他自己挖掘

✎ 我们给定了标签，限制了模型就干啥，就比如我就被限制要好好学习从而没能。。。

SimCLR Framework

- ✓ 我们小时候咋学习来着
 - ✎ 认识的信息很有限
 - ✎ 大部分负例都没见过
 - ✎ 通过对比来分析谁是谁
 - ✎ 其实这就是今天的故事了

Match the correct animal



SimCLR Framework

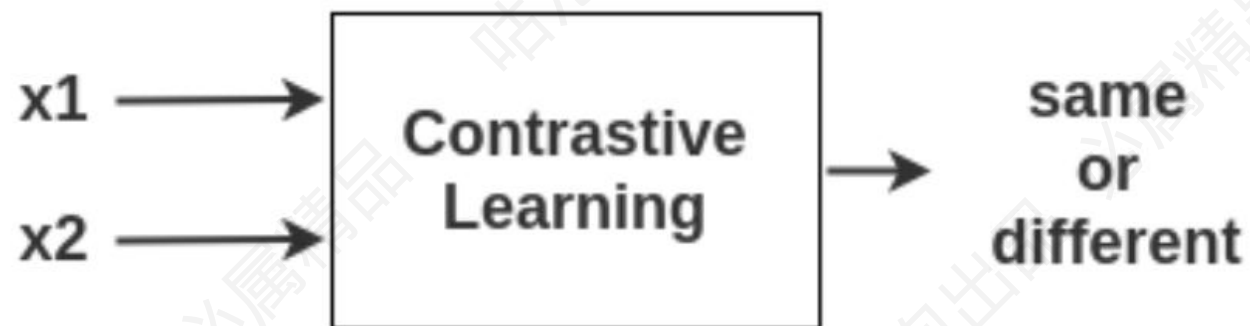
✓ 对比学习

✎ 其实就是判断异同

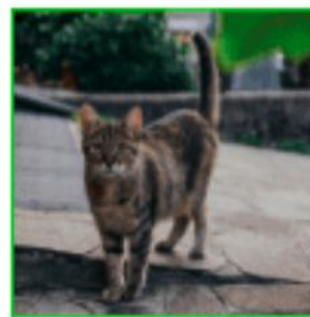
✎ 相同的就是正例

✎ 不同的就是负例

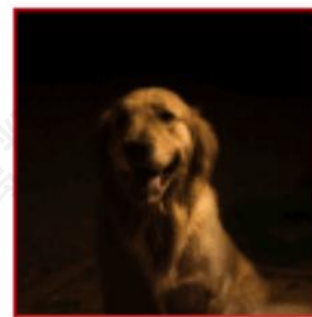
✎ 让模型学其中规律



Image



Similar



Different



Different

SimCLR Framework

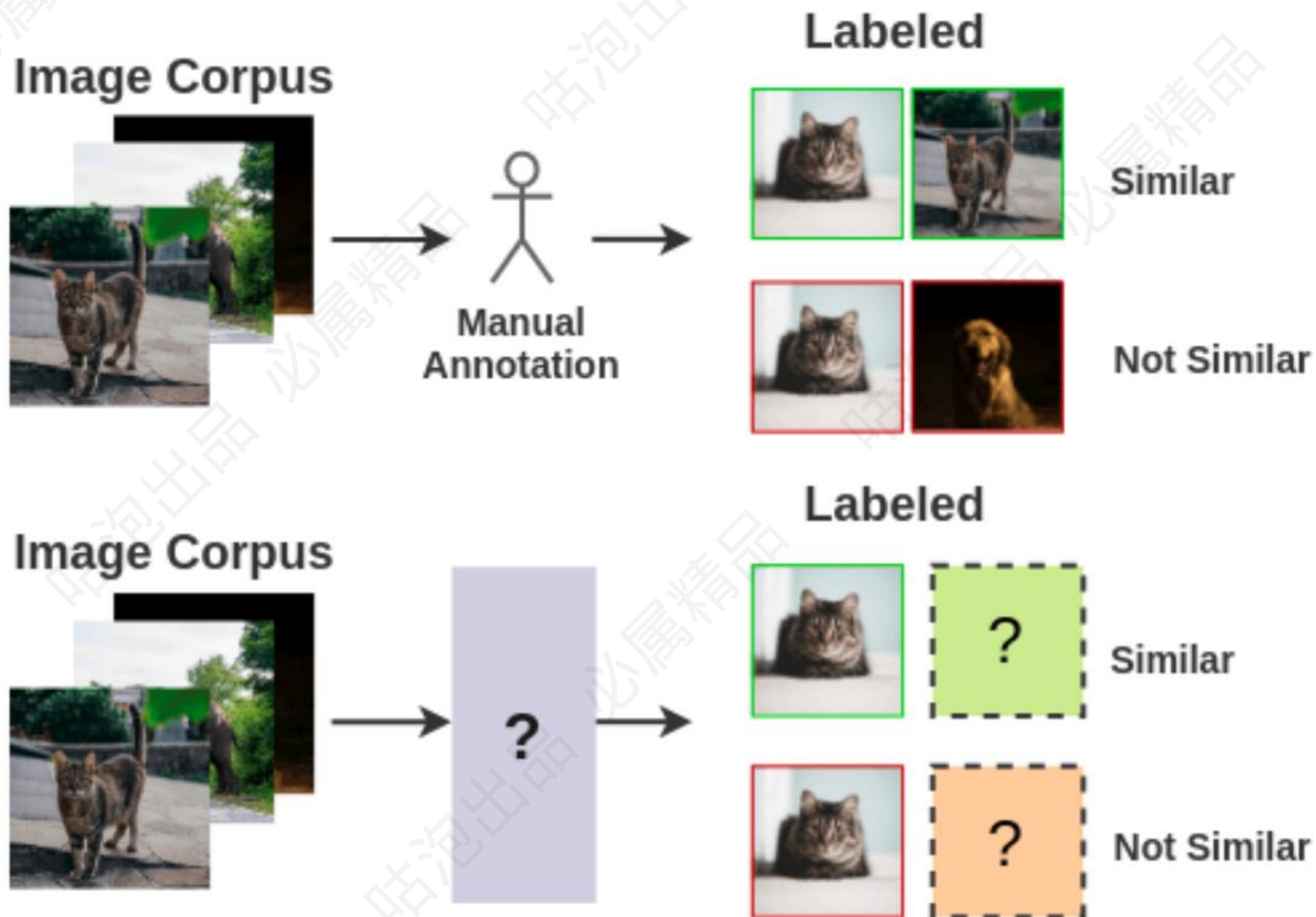
✓ 标签如何定义

✎ 要不要标注呢?

✎ 还得标注就没意义了

✎ 直接标下游任务就得了

✎ 还这么费劲搞什么对比



SimCLR Framework

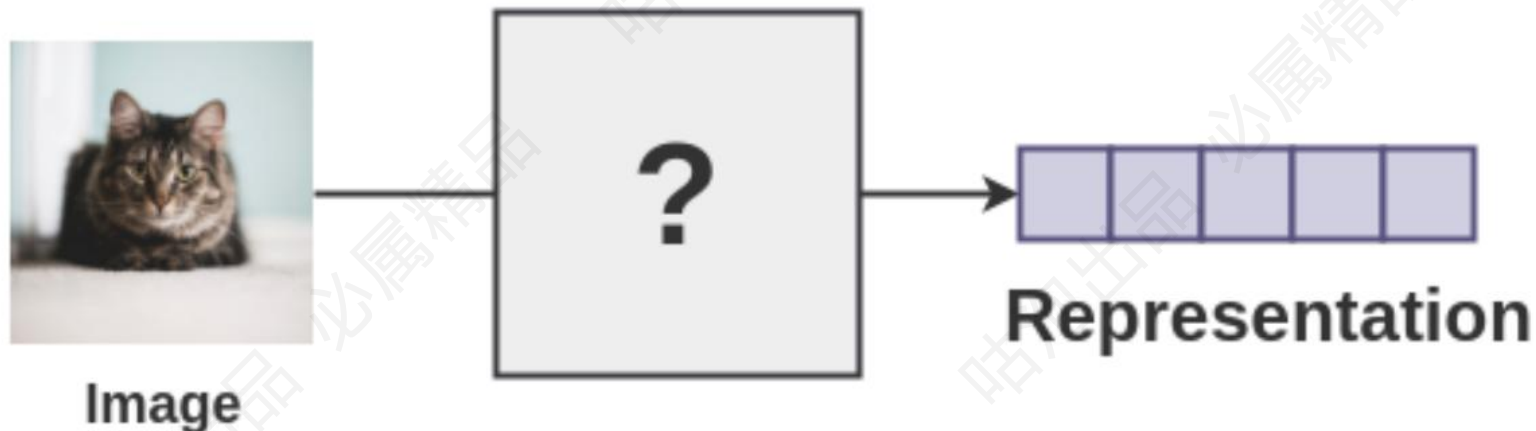
✓ 如何表示特征

✎ 图像最终得做成向量

✎ 这个简单，例如Resnet

✎ 还需要定义相似度的函数

✎ 来计算正负样本之间的距离

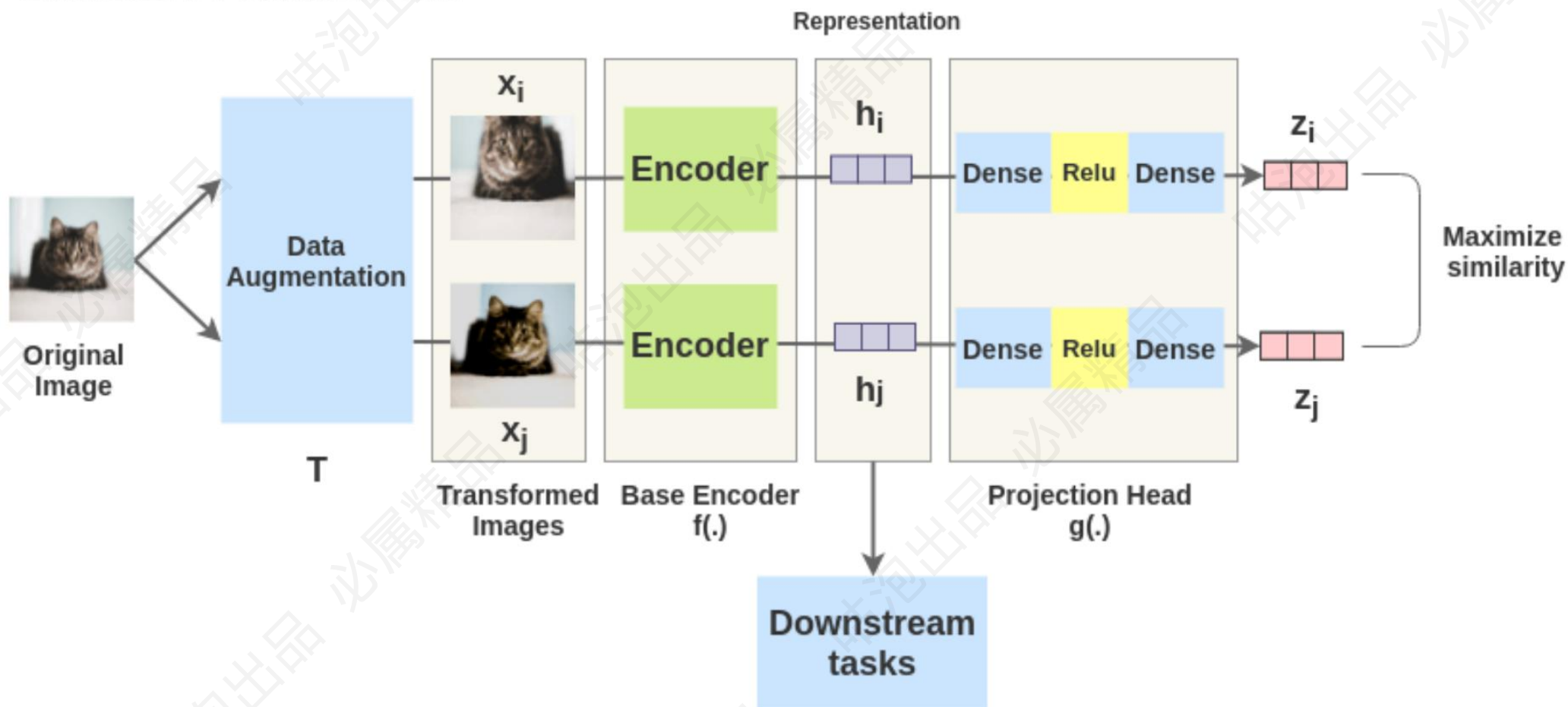


similarity(, )

SimCLR Framework

✓ 其实思想还是灰常简单的，不需要标签，自动就能学

SimCLR Framework



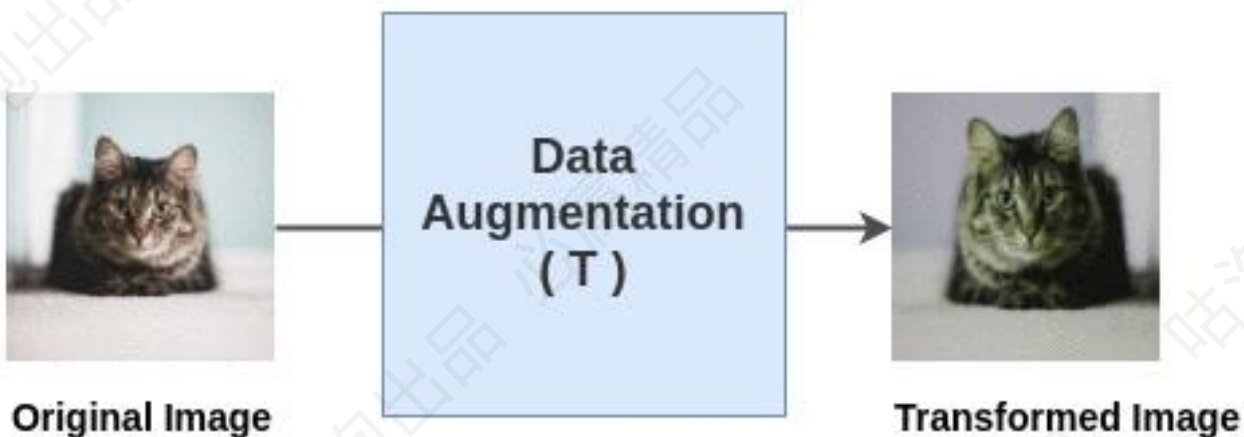
SimCLR Framework

✓ 有数据就足够了

✎ 估计大家也能猜到，要做正负样本了
(没有负样本就机器不学习了)

✎ 其实这里面BATCH比较重要，原文8192

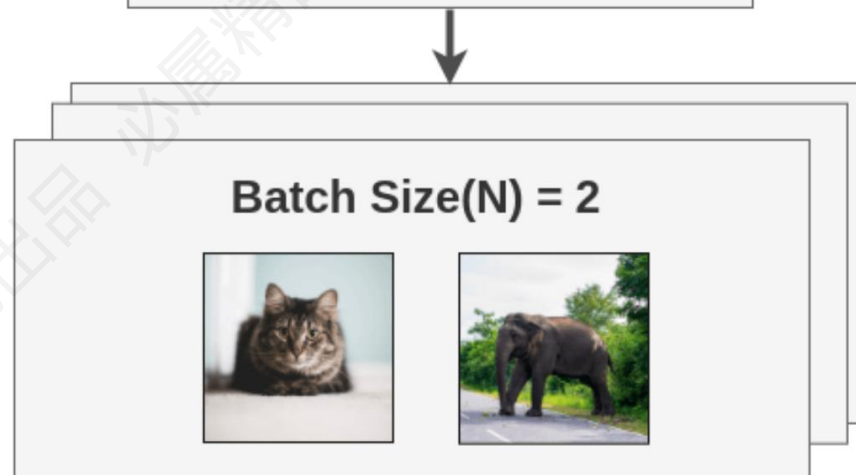
Random Transformation



Raw Corpus of Images

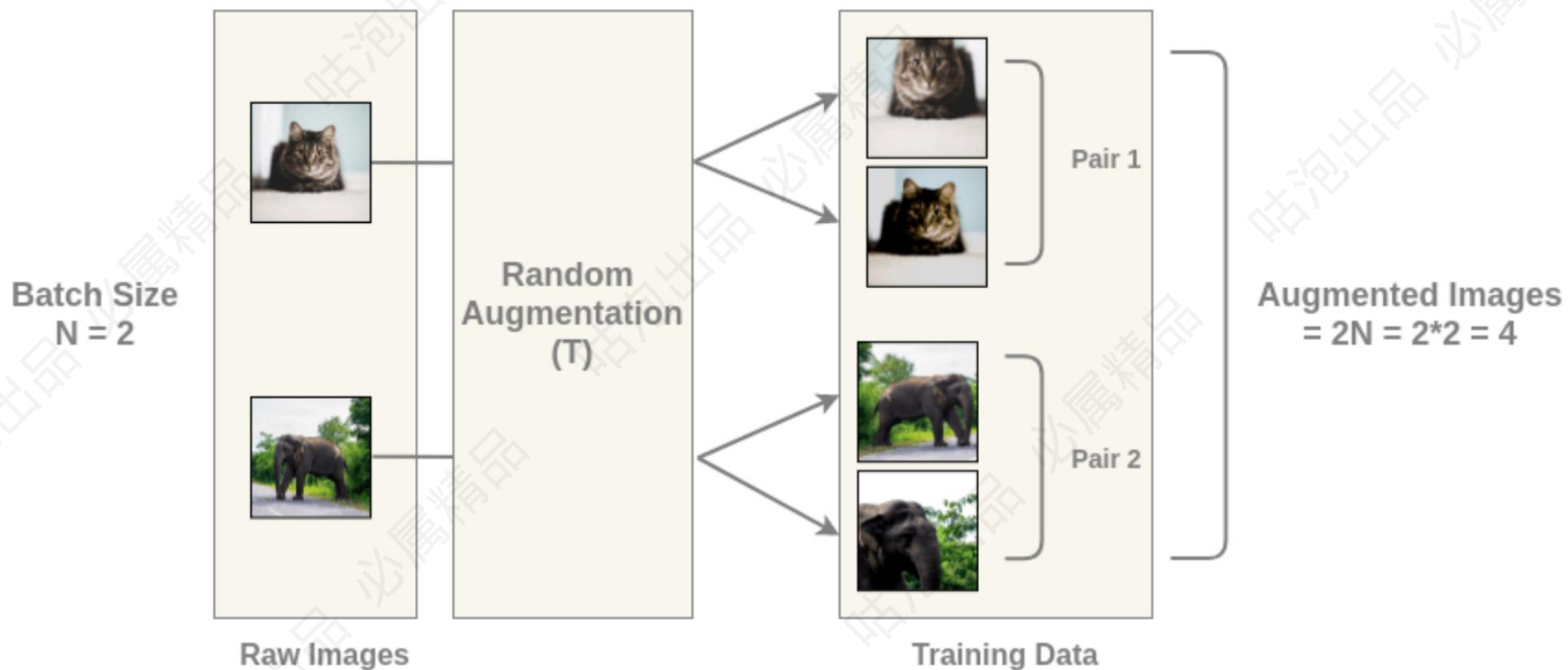


Whole Training Corpus



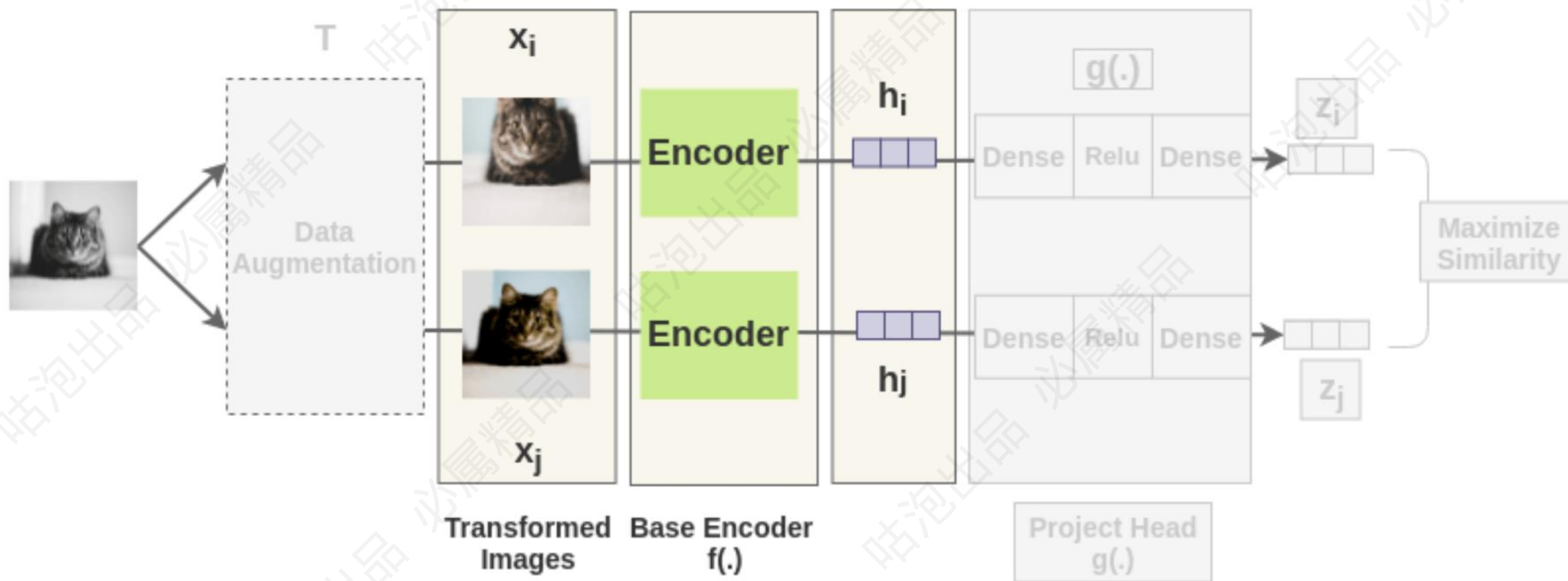
SimCLR Framework

✓ 数据增强得到咱们的一组输入



SimCLR Framework

✓ 特征也容易, ViT, resnet之类的都可以



SimCLR Framework

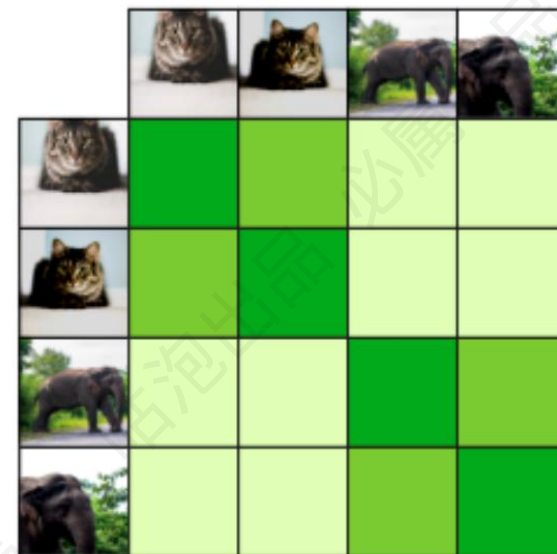
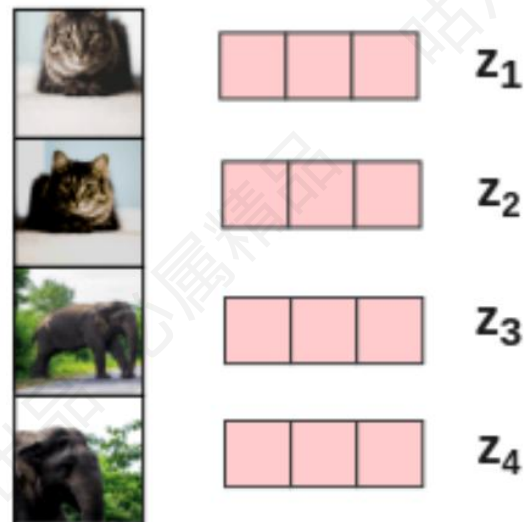
✓ 基本思想

✎ 其实就是同类越相似

✎ 一般用余弦相似度来定义

✎ 一般需要去掉对角线

✎ 但是损失函数怎么设计呢？

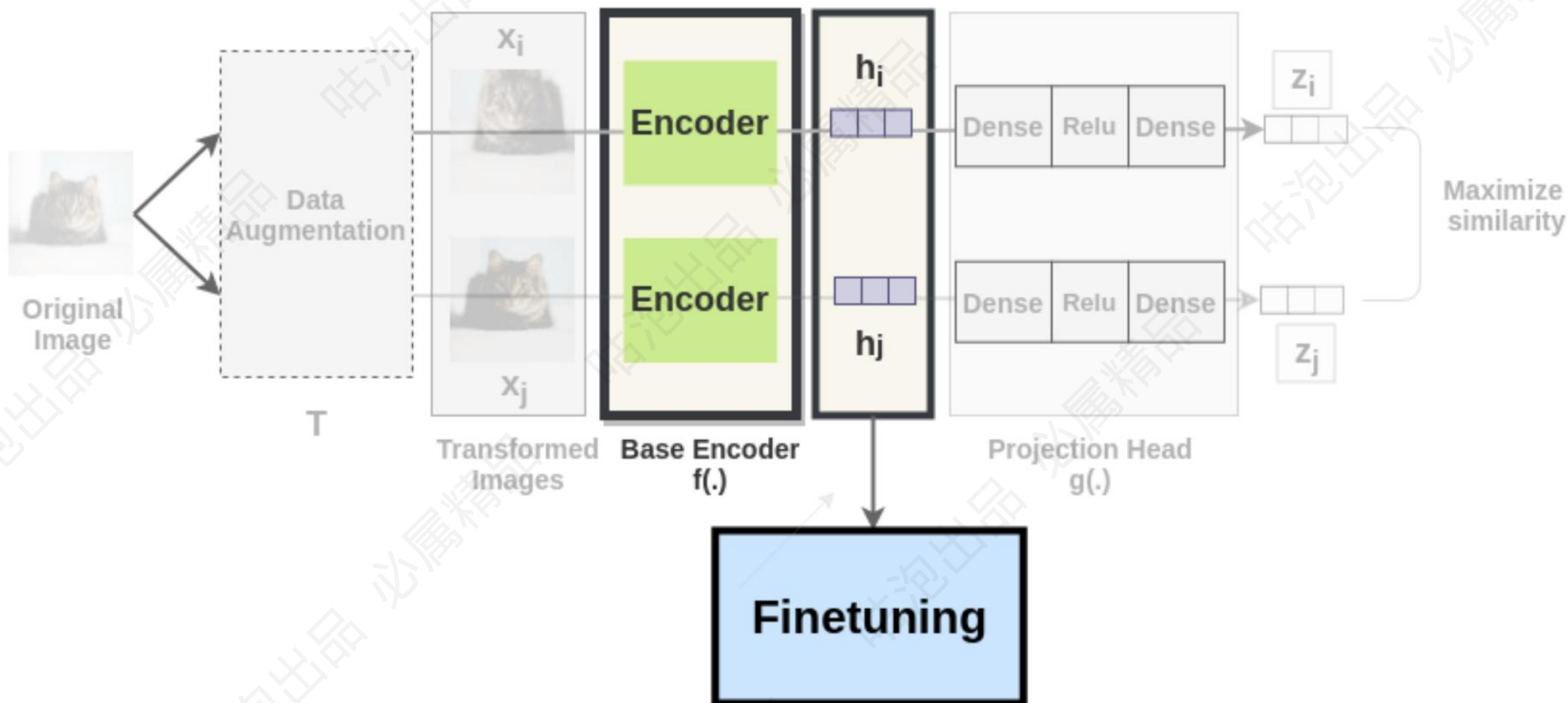


$$\text{similarity}(\mathbf{x}_i, \mathbf{x}_j) = \text{cosine similarity}(\mathbf{z}_i, \mathbf{z}_j)$$

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

SimCLR Framework

✓ 下游任务也很简单，直接用输出特征，注意不是MLP后的



SimCLR Framework

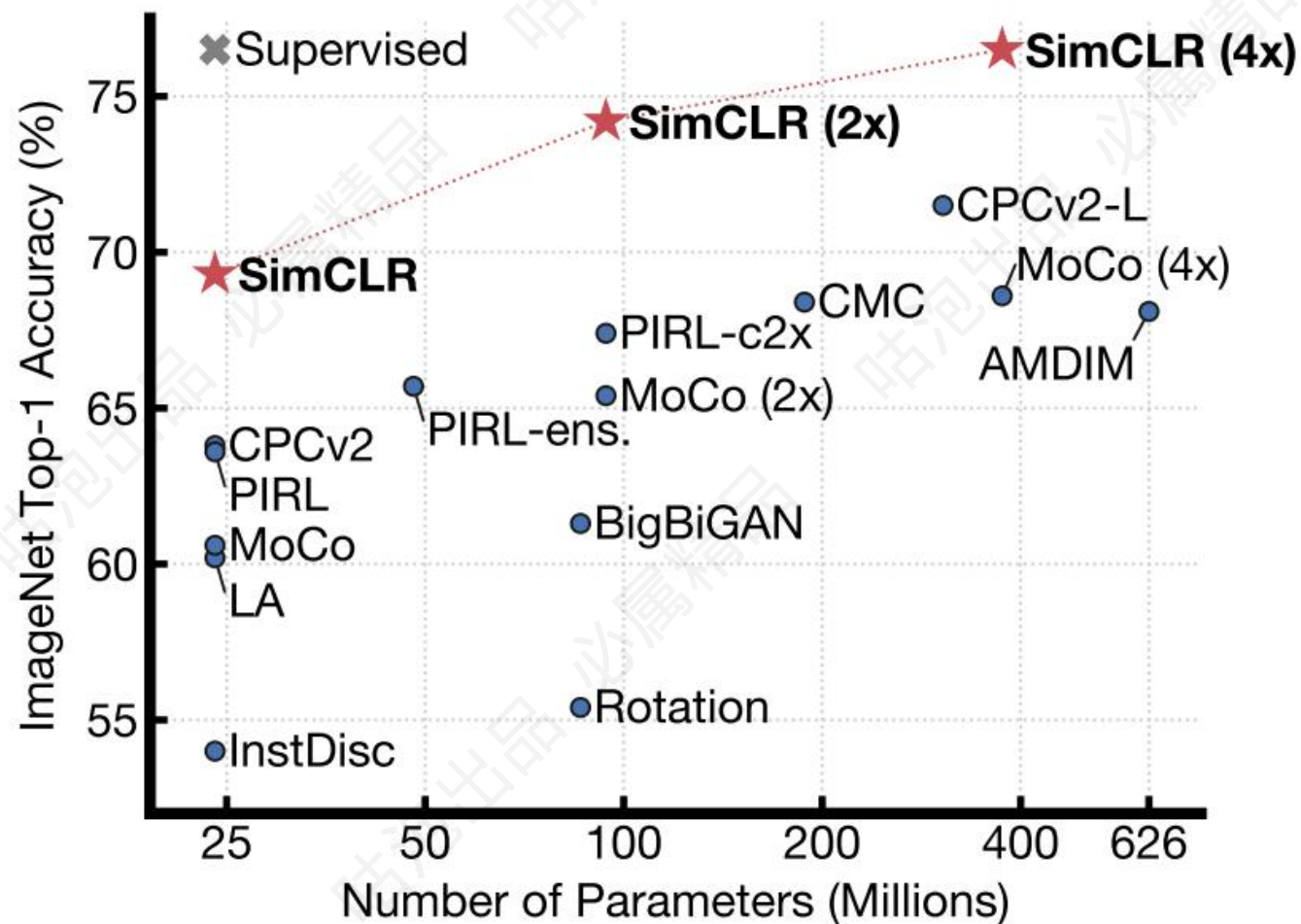
✓ 实验分析

✎ 虽然跟有监督还有距离

✎ 但是不需要标签啊

✎ 泛化能力肯定更强

✎ 视觉大模型啥时候现身



SimCLR Framework

✓ 数据处理中的变换

✎ 感觉还是得丰富一些，其实就是增大难度



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate {90°, 180°, 270°}



(g) Cutout



(h) Gaussian noise



(i) Gaussian blur



(j) Sobel filtering

SimCLR Framework

✓ 数据增强成为核心了

✎ 越花了呼哨没准越好

✎ 越离谱需要学的越多

✎ 相同的就很低，简单

✎ Crop和Color的比较配



SimCLR Framework

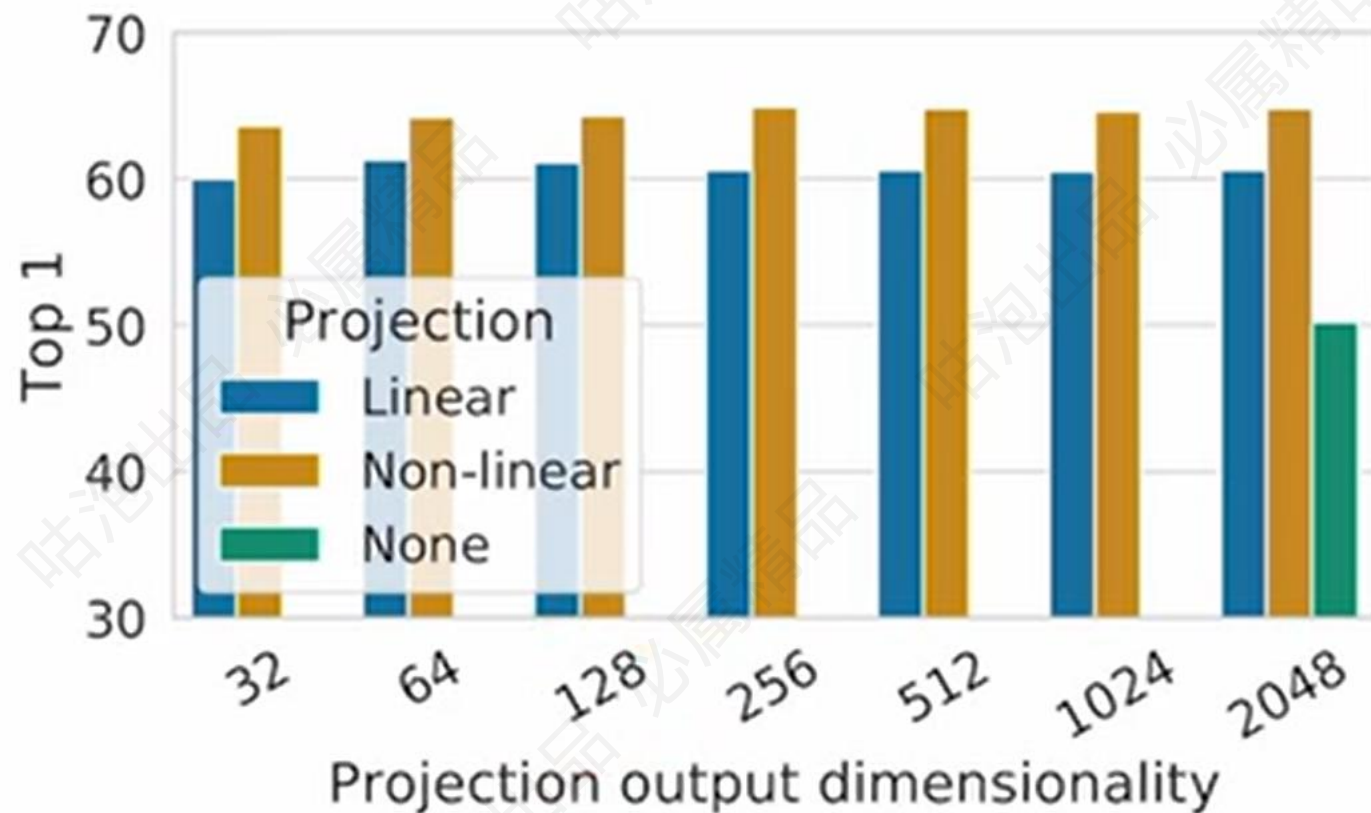
✓ 最后的MLP

✎ 这东西还给吹了吹

✎ 反正就是加上好

✎ 但是维度感觉有点奇怪

✎ 32与2048平起平坐了



SimCLR Framework

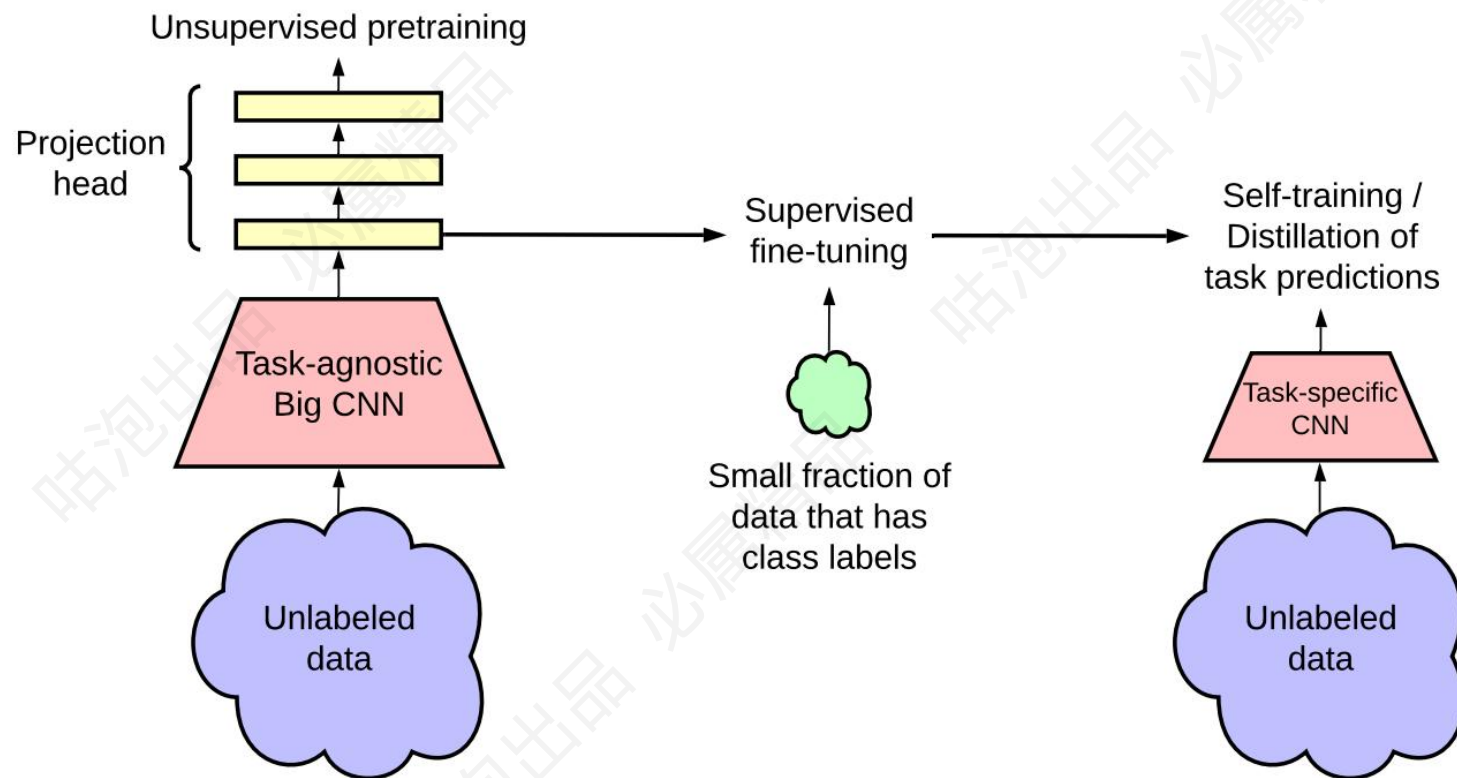
✓ V2版本其实就一个事。。。

✎ 模型做的更大了

✎ 然后好像没啥特别的了

✎ 又加了一个蒸馏

✎ 效果强了一些，就完了



Multiview Coding

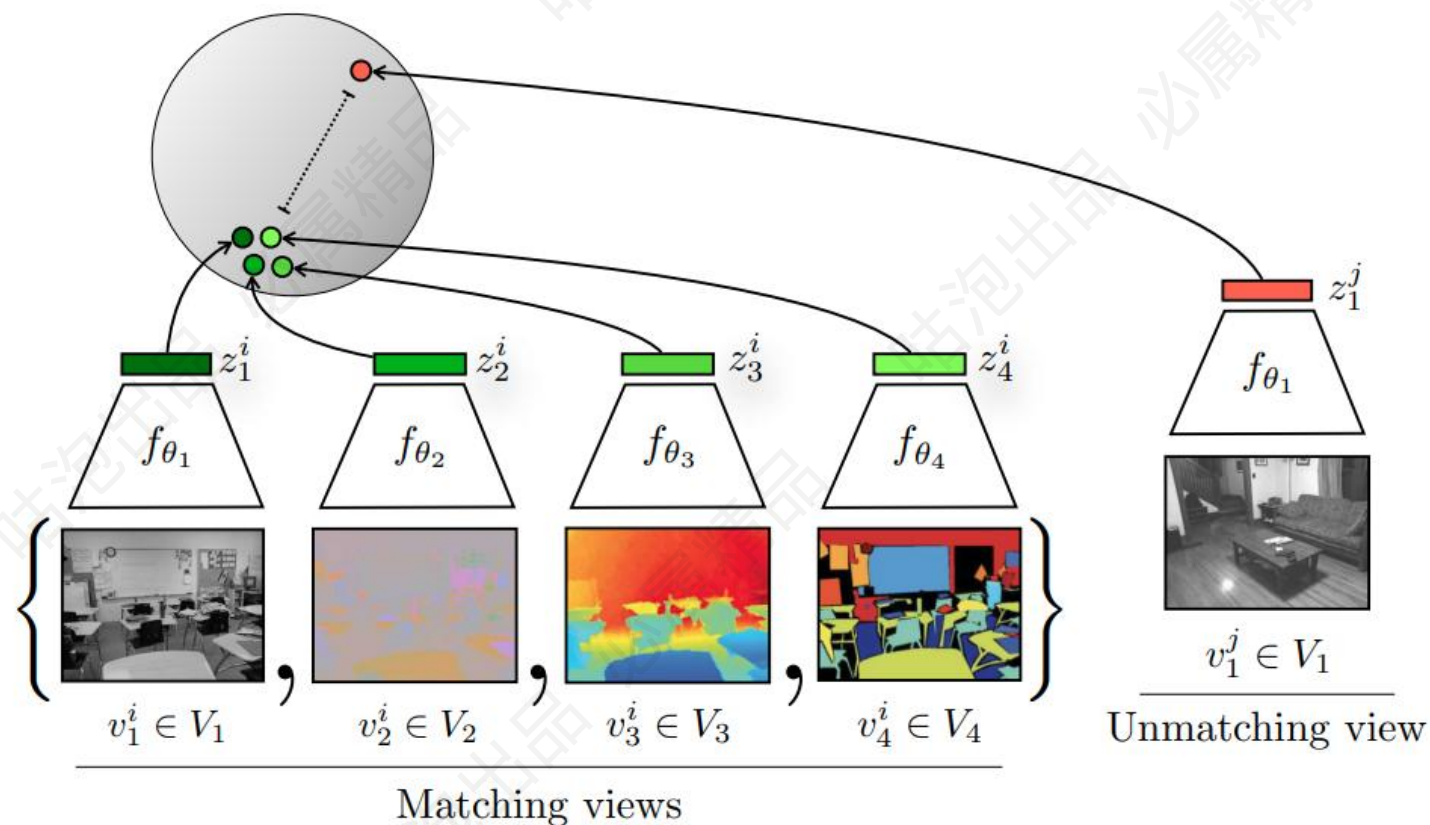
✓ 多视角任务

✎ 你就算化成灰我都认识

✎ 这才叫真正的理解。。。

✎ 不同视角的特征都是同类

✎ 分割的，深度的等等

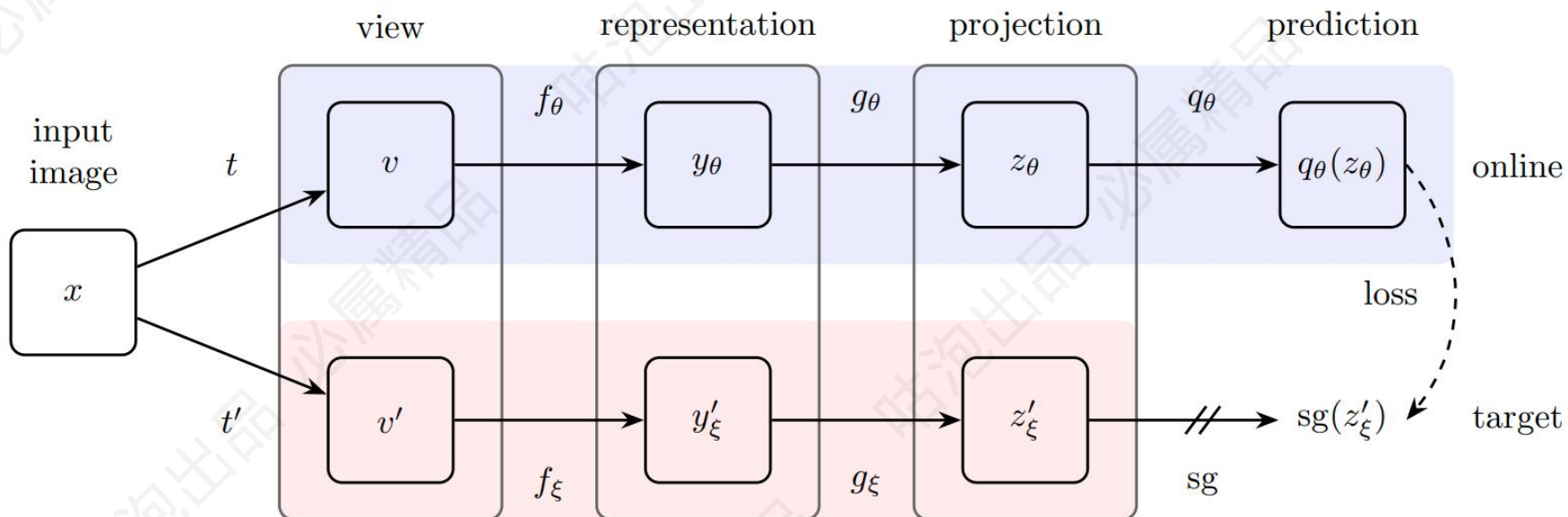


BYOL

✓ 这是离了个大谱

✎ 咱们之前一顿吹负样本咋咋滴的，增加难度，训练能好之类的

✎ 现在你告诉我，不需要负样本这个事也能办了？

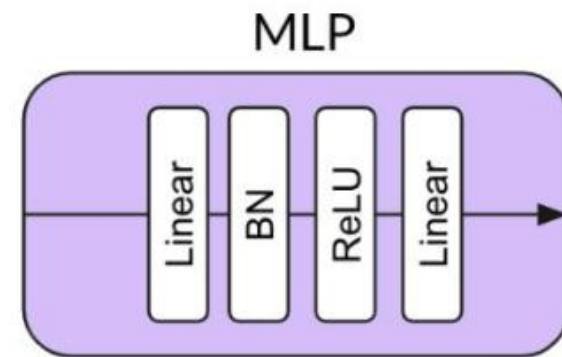
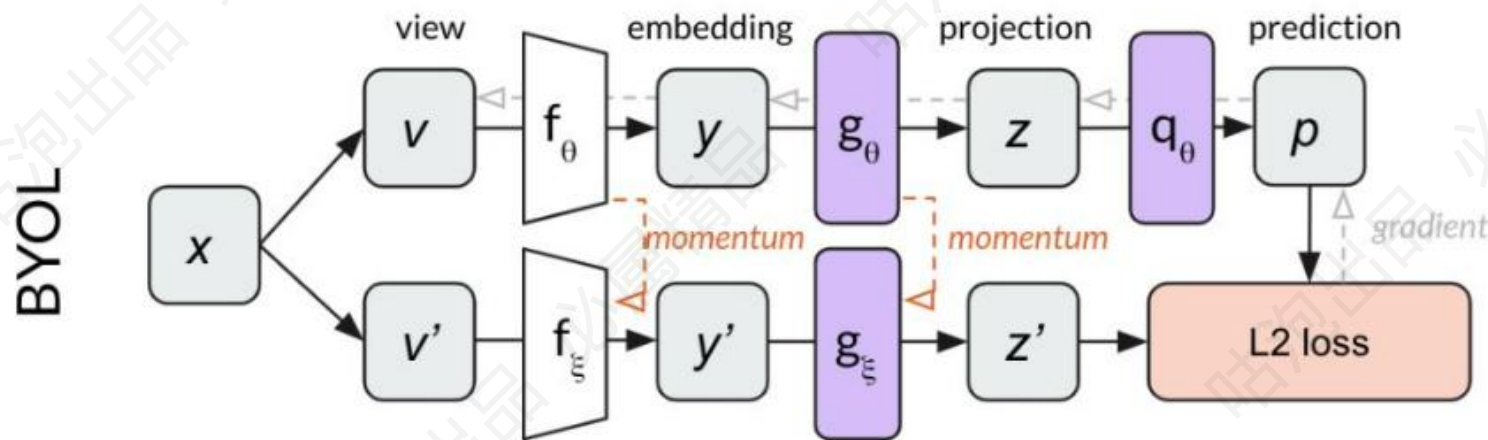


BYOL

✓ 这是离了个大谱

✎ 你说这个事该怎么解释呢，有大佬指出“罪魁祸首”竟然是BN

✎ BN啥意思来着，算这一批样本的，等等你说啥，这一批？那不就包括了负样本



SimCSE Framework

✓ 如何提取句子向量

✎ 句子向量如何获取呢？我估计大家第一个就能想到BERT

✎ 大概率是直接取CLS的向量了，但是这样做会不会有啥问题呢？

✎ BERT训练的是分类任务，但是分类会不会限制住模型的能力呢？

✎ 这回咱们换换套路，用对比学习的方法来讲NLP句子的故事

SimCSE Framework

✓ 文本任务如何提特征

✎ 这啥说这篇论文呢，方法贼简单，但是效果还能挺好

✎ 文本如何套对比学习呢？怎么定义正负样本呢？

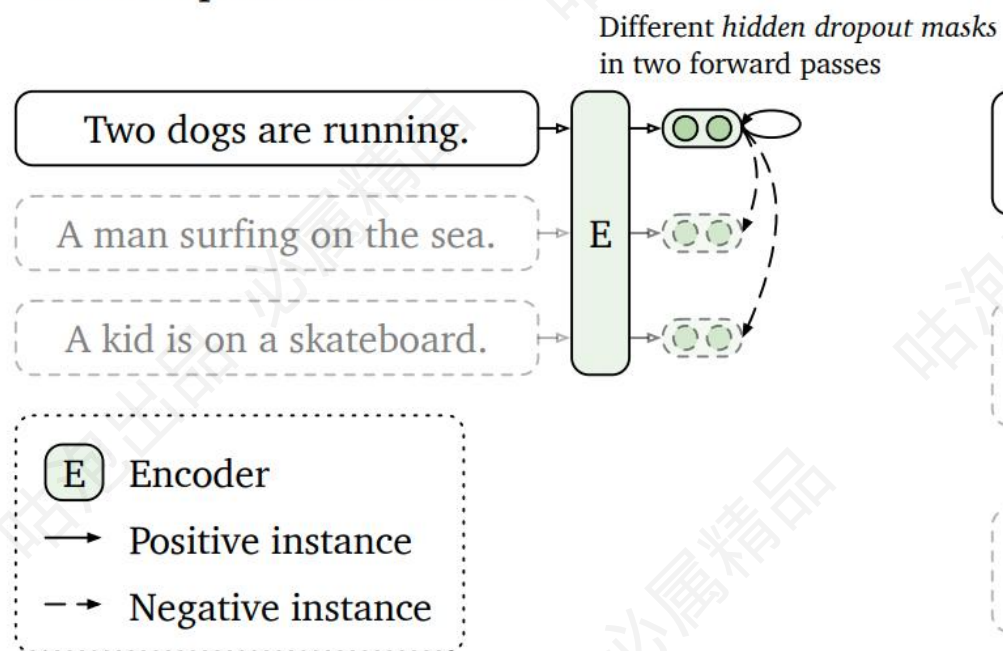
✎ 1.我今天打DOTA暴走了； 2.我今天打英雄联盟拿了五杀

✎ 1.我喜欢打DOTA； 2.我不喜欢打DOTA

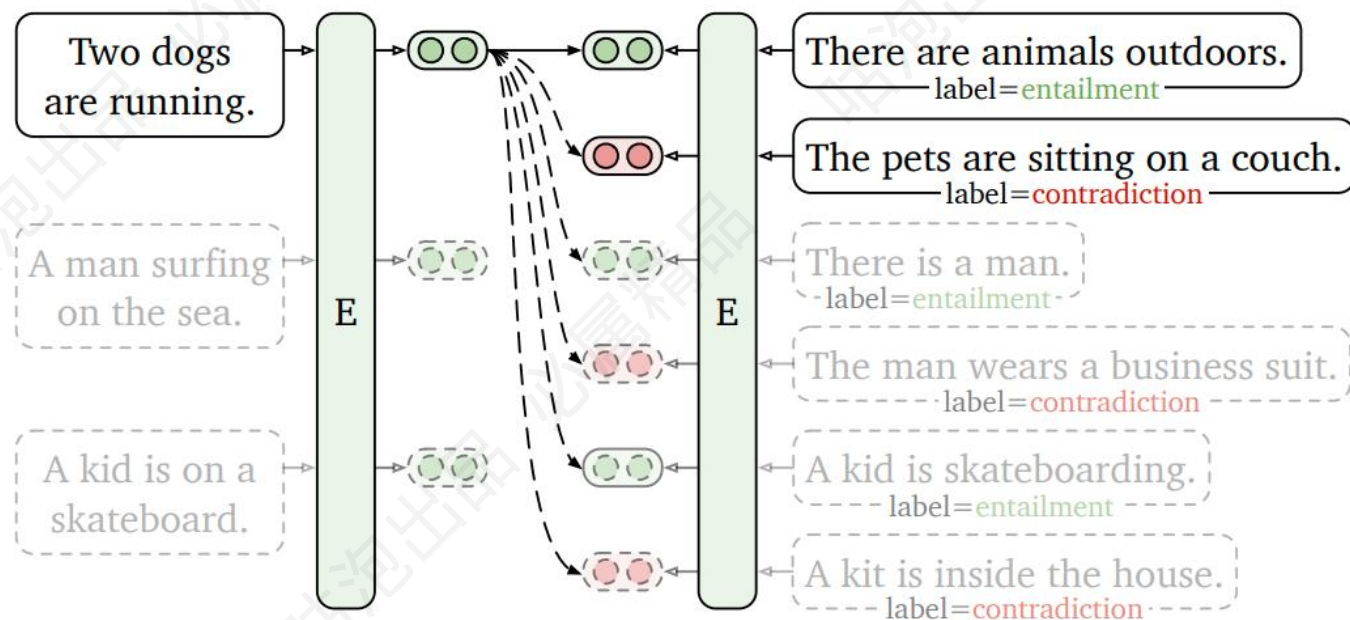
SimCSE Framework

✓ Dropout成功超神

(a) Unsupervised SimCSE



(b) Supervised SimCSE



SimCSE Framework

✓ 评估分析

✎ 就做了这么简单点事

✎ 却让结果显著提升了。。

✎ 句子提特征也有招了

✎ align表示同类之间的距离
Uniform表示所有句子整体分布

