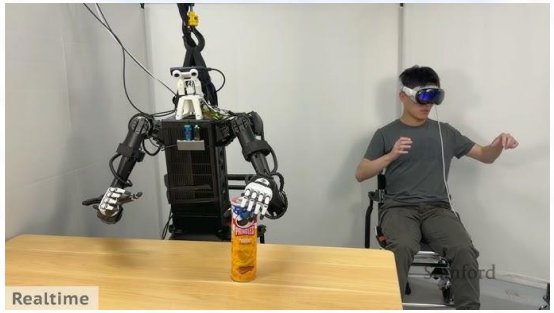
Self-Supervised Cross-View Correspondence with Predictive Cycle Consistency

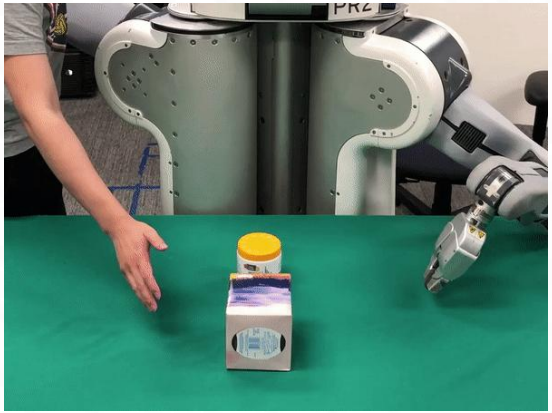
汇报人：张淋迩

2025.11.3

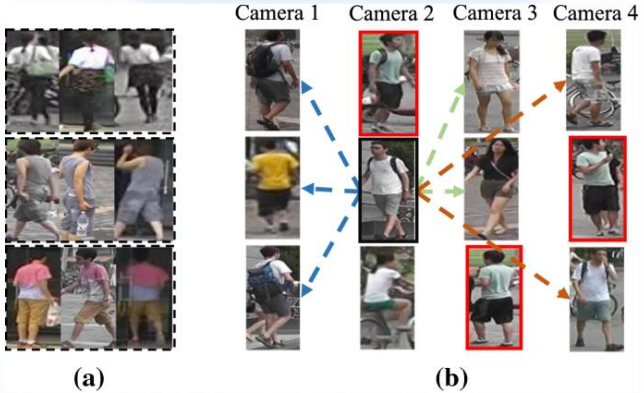
01 研究背景与问题

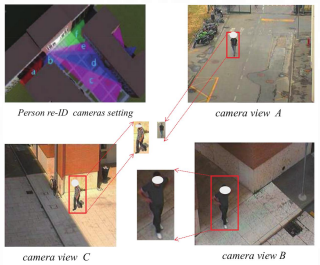
机器人模仿学习

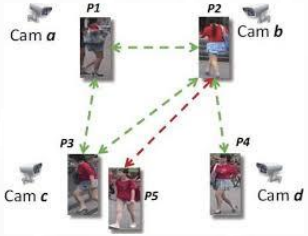




目标重识别







视频目标跟踪

Cross-View Correspondence

传统监督学习阶段 2018年前

核心依赖人工标注的像素级 / 目标级对应关系，代表方法如 SIFT、SURF 等手工特征匹配算法，以及基于深度学习的 FlowNet、XSegTx。这类方法虽精度较高，但存在两大致命缺陷：

- 标注成本极高
- 泛化性不足

自监督初步探索阶段 2018-2022年

以“代理任务 + 无标注数据”为核心，通过图像自身的内在一致性构建监督信号，主要分为三大技术路线：

- 光度一致性重建路线（MVSNet 系列方法）
- 对比学习路线（CrossPoint）
- 循环一致性路线（Cycle-Consistency）

自监督进阶优化阶段 2022-2024年

- 语义增强优化：该方法需依赖多视图的固定场景假设，且协同分割的精度受限于预训练 CNN 特征
- 长时序适应优化：这类方法需依赖 RGB-D 的深度信息，而真实场景中深度传感器并非总能获取（LVOS）
- 掩码预测优化：该方法的预测表征受语义干扰强，在语义相似目标的对应性匹配中易混淆（SiamMAE）

极端视角适应能力缺失

have overwhelmingly focused on domains with continuous or small transformations, such as Video Object Segmentation (VOS) in continuous videos, or dense pixel-to-pixel methods which necessarily assume that object surfaces overlap between views. This limits the application of self-supervised correspondence to *discontinuous* inputs,

长时时空一致性建模不足

The fundamental challenge of learning visual correspondence across discontinuities is that a system must take into account an object and the surrounding environment holistically to answer the question of where *that* mug is instead of where *a* mug is. A particularly powerful approach for en-

自监督学习三大瓶颈

语义与空间信息的表征纠缠

scenarios (Table 1). State-of-the-art approaches typically rely on K-Nearest-Neighbor matching of emergent learned representations, identifying corresponding regions by comparing the most similar model embeddings across views. Although the pretraining objectives of these methods promote spatial awareness within a scene, the resulting representations often entangle semantic object information with the spatial cues necessary for correspondence. This entanglement reduces robustness against semantically similar distractor objects, limiting generalization and performance in more demanding scenarios.

Cross-View Correspondence

三大核心贡献 实现针对性突破

1. We introduce Predictive Cycle Consistency, a technique that combines the powerful representation learning of predictive approaches with the refinement of cycle-consistency for self-supervised correspondence.
2. We propose a pseudo-labeling method that incorporates cycle-consistency on top of existing correspondence models to iteratively refine self-supervised object correspondence outputs.
3. We obtain state-of-the-art results on a suite of correspondence tasks from EgoExo4D, DAVIS, and LVOS.

提出预测性循环一致性框架（PCC），首次融合“预测学习”与“循环一致性”优势。
PCC 通过“条件灰度着色预测”自动生成初始对应关系，解决“极端视角适应缺失”问题。

设计迭代伪标签优化策略，攻克“长时时空一致性建模不足”难题。
针对 SiamMAE 等方法“长间隔性能骤降”的缺陷，提出“伪标签生成 - 模型训练 - 伪标签迭代优化”的闭环流程。

Method	Backbone	Ego Query				Exo Query			
		Bal. Acc. ↑	IoU ↑	Loc. Score ↓	CA.↑	Bal. Acc.↑	IoU↑	Loc. Score↓	CA.↑
XSegTx w/o finetuning [16]	SegSwap [45] + ViT-B	∅	0.60	0.116	0.017	∅	1.62	0.197	0.027
XSegTx [16]	SegSwap [45] + ViT-B	62.63	13.88	0.154	0.239	74.6	21.8	0.133	0.265
XMem♣ [6]	ResNet-50 [20] + Memory	42.33	13.07	0.312	0.182	56.96	10.2	0.249	0.125
XView-XMem♣ [16]	XMem + ViT-B	53.28	22.14	0.176	0.325	59.36	23.56	0.186	0.308
XView-XMem (+ XSegTx)♣ [16]	XMem + SegSwa + ViT-B	54.61	22.5	0.139	0.347	52.28	19.39	0.208	0.255
Ours Supervised	ViT-B/16	74.7	38.41	0.037	**0.603**	**88.45**	43.70	**0.049**	0.555
Ours Supervised + PCC	ViT-B/16	**76.9**	**39.01**	**0.033**	0.600	87.23	**47.06**	0.054	**0.590**
SiamMAE◇ [18]	ViT-S/8	∅	12.24	0.170	0.185	∅	14.18	0.159	0.198
CrocoV2◇ [63]	ViT-B/16	∅	7.14	0.200	0.138	∅	9.56	0.164	0.136
DINO [4]	ViT-B/8	∅	12.55	0.153	0.178	∅	15.94	0.137	0.246
DINOv2+Registers [37] [8]	ViT-B/14	∅	20.26	0.125	0.299	∅	24.6	0.169	0.307
SiamMAE◇+SAM	ViT-S/8+ViT-H	∅	17.97	0.180	0.254	∅	24.05	0.143	0.315
DINOv2+Registers+SAM	ViT-B/14+ViT-H	∅	28.92	0.153	0.365	∅	34.78	0.123	0.433
Grayscale Coloration + SAM◇	ViT-B/16+ViT-H	∅	20.82	0.110	0.311	∅	19.50	0.109	0.276
PCC Iter 1	ViT-B/16	∅	26.41	0.085	0.396	∅	34.35	0.090	0.436
PCC Iter 2	ViT-B/16	65.22	**29.98**	**0.083**	**0.446**	66.40	40.41	0.079	0.502
PCC Iter 3	ViT-B/16	60.66	29.89	0.094	0.432	67.90	**41.45**	**0.071**	**0.508**

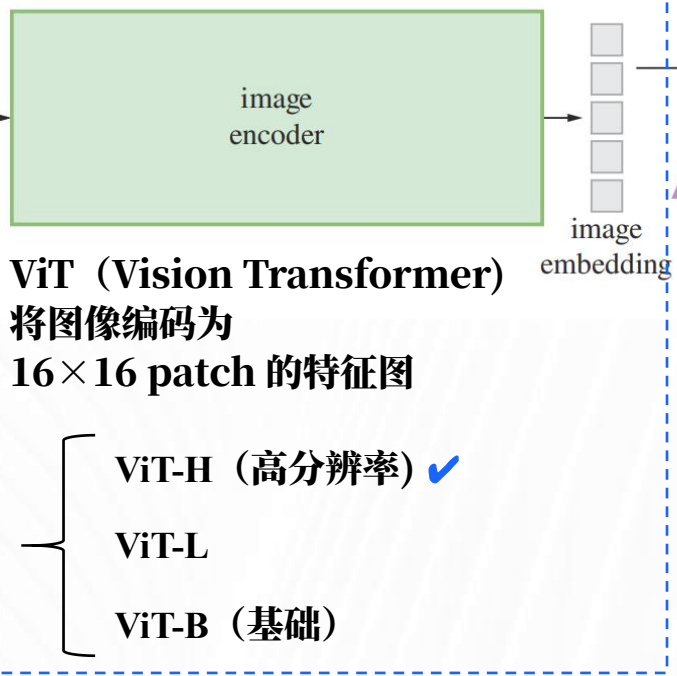
在EgoExo4D、davis和LVOS等数据集的对应任务中取得了最先进的成果

02 核心方法

SAM (Segment Anything Model)

- Meta AI (2023)
- 打破传统分割模型“需场景特定标注”的限制，实现“零标注 / 少标注下精准分割任意目标”，覆盖物体、背景、小目标、遮挡目标等各类场景

image



ViT（Vision Transformer）将图像编码为 16×16 patch 的特征图

- ViT-H（高分辨率）✔
- ViT-L
- ViT-B（基础）

输出：

$O_{1,i}$ 是 I_1 图像中分割出的物体，$i \in N_1$

$O_{2,j}$ 是 I_2 图像中分割出的物体，$j \in N_2$

points（点）：用户标出目标或背景点

box（框）：用户画出矩形框

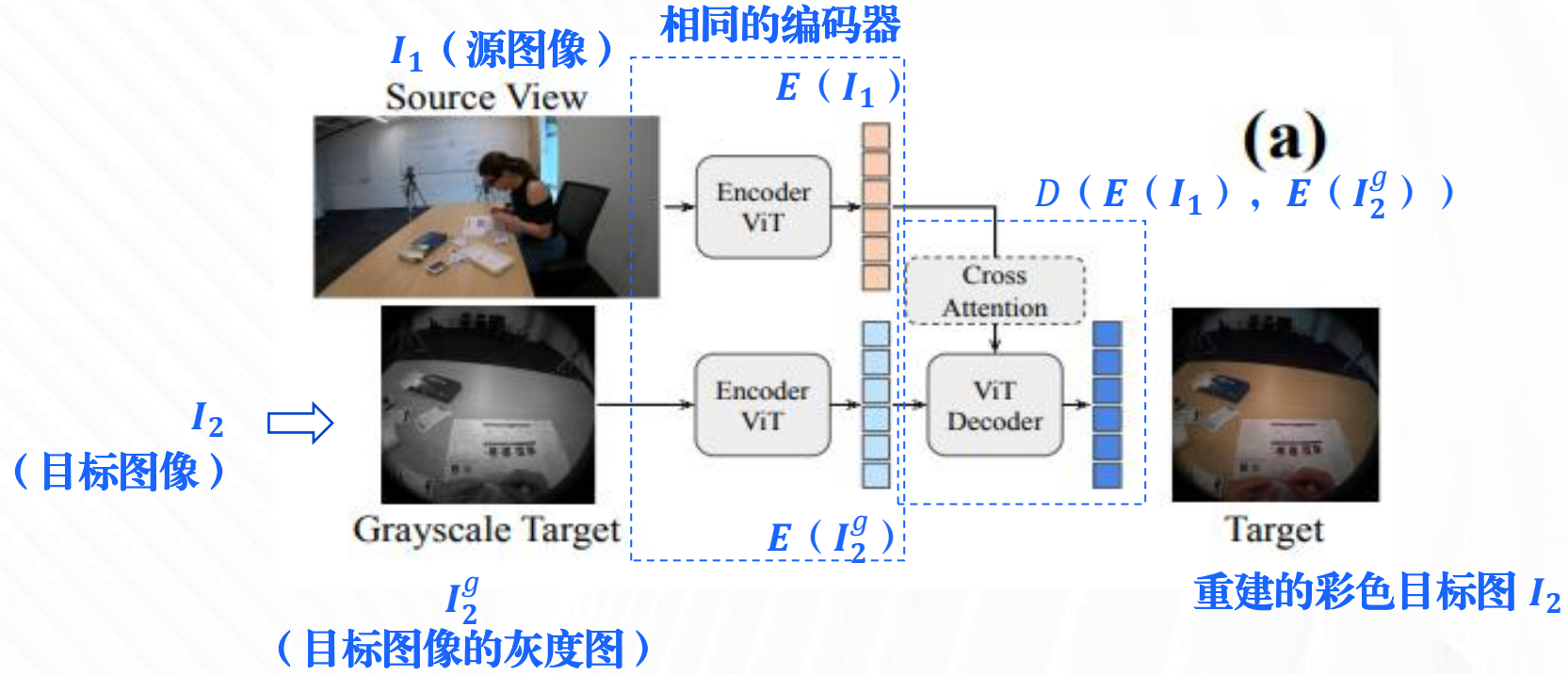
text（文本）：未来可能扩展到文本描述

mask（掩膜）：已有掩膜信息

PCC 中启用“自动分割模式”（auto_mask_generator）

基于灰度着色的热力图提取

- 利用“色彩一致性”推断物体对应：若 I_1 中的物体 $O_{1,i}$ 与 I_2 中的物体 $O_{2,j}$ 是同一物体，则给 $O_{1,i}$ 加色后，$O_{2,j}$ 在 I_2 的重建彩色图中颜色也会变化

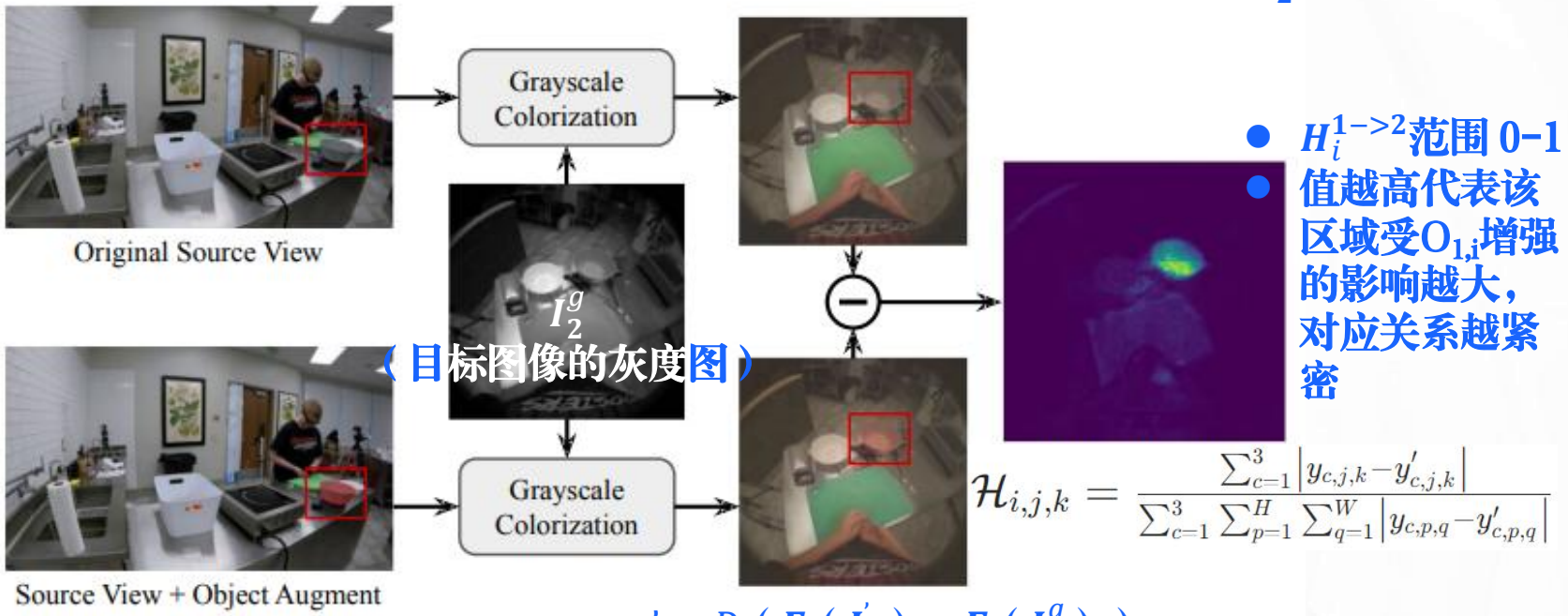


采用**RGB 空间的均方误差（MSE）损失**，公式为：$\mathcal{L}_{MSE} = \frac{1}{3HW}\sum_{c=1}^{3}\sum_{i=1}^{H}\sum_{j=1}^{W}(y_{c,i,j} - I_{2,c,i,j})^2$

基于灰度着色的热力图提取

- 利用“色彩一致性”推断物体对应：若 I_1 中的物体 $O_{1,i}$ 与 I_2 中的物体 $O_{2,j}$ 是同一物体，则给 $O_{1,i}$ 加色后，$O_{2,j}$ 在 I_2 的重建彩色图中颜色也会变化

$$y = D(E(I_1), E(I_2^g))$$



I_1（源图像）

I_2^g（目标图像的灰度图）

$I_1^{'}$（对$O_{1,i}$进行“色彩增强”即给$O_{1,i}$区域的 RGB 三通道各加一个固定偏移c=[20,20,20]）

$$y' = D(E(I_1^{'}), E(I_2^g))$$

- $H_i^{1->2}$范围 0-1 值越高代表该区域受$O_{1,i}$增强的影响越大，对应关系越紧密

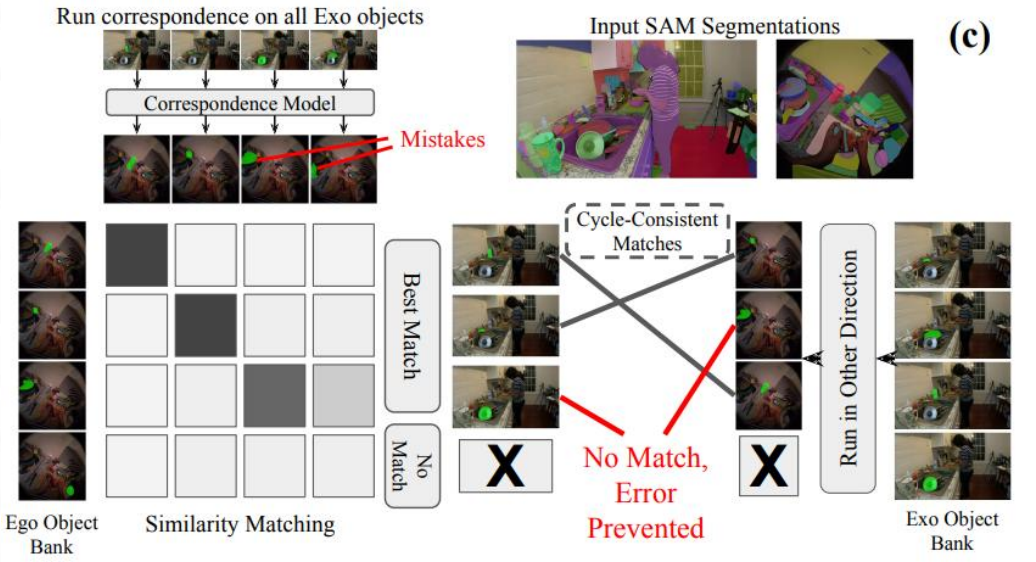
$$\mathcal{H}_{i,j,k} = \frac{\sum_{c=1}^{3} |y_{c,j,k} - y'_{c,j,k}|}{\sum_{c=1}^{3}\sum_{p=1}^{H}\sum_{q=1}^{W} |y_{c,p,q} - y'_{c,p,q}|}$$

循环一致性筛选

- 仅通过“1→2 方向”的热度图筛选对应对，易出现“单向错误”（如$O_{1,i}$的热度图最高区域对应$O_{2,j}$，但$O_{2,j}$的热度图最高区域不对应$O_{1,i}$）

⇩ 论文引入“循环一致性”进行双向验证

- 交换I_1和I_2的角色，重复前述步骤，得到反向热度图$H_j^{2->1}$（$O_{2,j}$在I_1中对应的热力图）



- 正向匹配

$$P_j^{1\to 2} = argmax_i Sim(H_i^{1\to 2}, O_{2,j})$$

- 反向匹配

$$P_i^{2\to 1} = argmax_j Sim(H_j^{2\to 1}, O_{1,i})$$

其中，相似度计算

$$Sim(A,B) = \sum_{j=1}^{H}\sum_{k=1}^{W} A_{j,k} \times B_{j,k}$$

正向与反向同时匹配时作为最终的伪标签

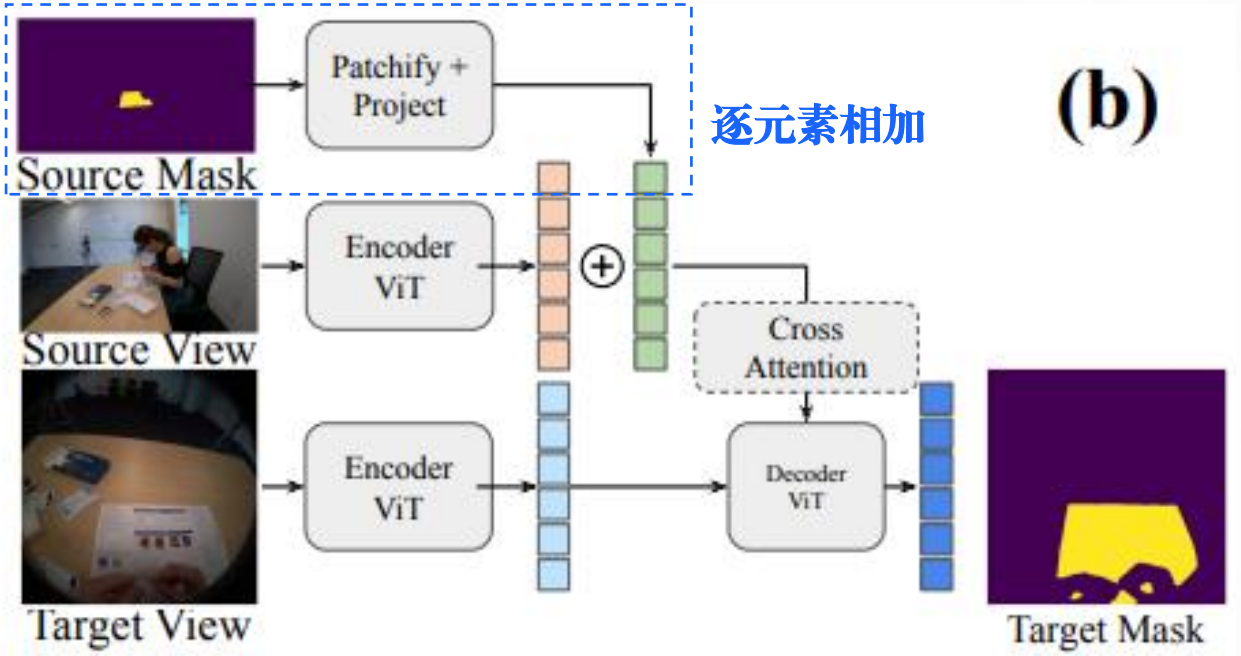
对应模型训练

- 前面步骤筛选出的“$O_{1,i}\leftrightarrow O_{2,j}$”对应关系，仅停留在“静态标注”层面，无法直接用于“给定一张图的物体掩码，实时预测另一张图对应掩码”的实际场景

O_1（源图像掩码）

I_1（源图像）

I_2（目标图像）



逐元素相加

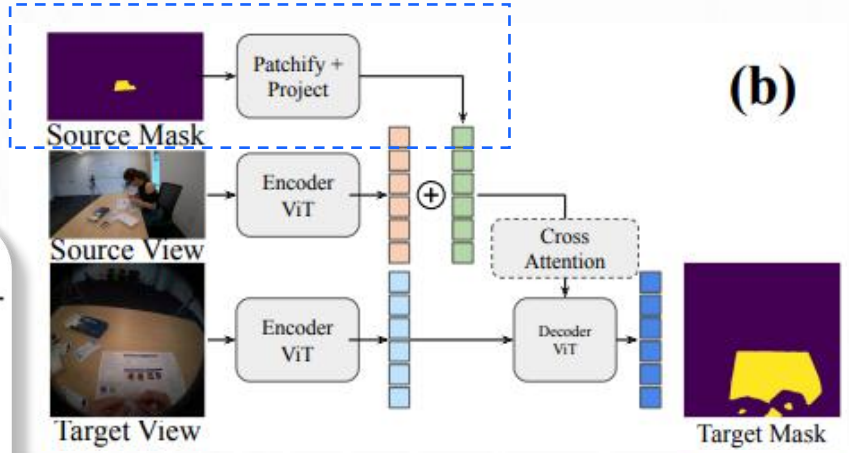
O_2（目标图像掩码）

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{j,k} O_{2,j,k} \times \hat{O}_{2,j,k}}{\sum_{j,k} O_{2,j,k}^2 + \sum_{j,k} \hat{O}_{2,j,k}^2} + \mathcal{L}_{BCE} = -\frac{1}{HW}\sum_{j,k}[O_{2,j,k}\log\hat{O}_{2,j,k} + (1 - O_{2,j,k})\log(1 - \hat{O}_{2,j,k})]$$

迭代优化

- 初始伪标签由“灰度着色模型”生成，精度有限
  - 用上一步骤练好的“对应模型”替换“灰度着色模型”，重新生成新热度图、筛选循环一致对，得到精度更高的伪标签

每一轮的伪标签更新
输入的源掩码也随之更新



对应增强的源图像的处理

- 相似度计算改为交并比IOU
- 更精准衡量“预测掩码”与“物体掩码”的重合度

Method	$\mathcal{J}\&\mathcal{F}_m\uparrow$	$\mathcal{J}_m\uparrow$	$\mathcal{F}_m$
SiamMAE [18]	60.7	58.4	62.9
CrocoV2 + Cont. Pretrain [63]	40.0	37.4	42.5
Probalistic Warp Consistency [54]	42.9	42.6	42.7
DINO ViTs/8 [4]	64.5	61.6	67.5
DINO ViTb/8 [4]	66.4	63.7	69.2
DINOv2 + Reg ViTb/14 [8]	62.1	59.6	64.8
PCC Iter 1	64.4	61.3	67.5
PCC Iter 2	69.7	67.0	72.4
PCC Iter 3	**70.2**	**67.8**	**72.7**

实验表明一次初始伪标签
和两次优化迭代后性能饱和

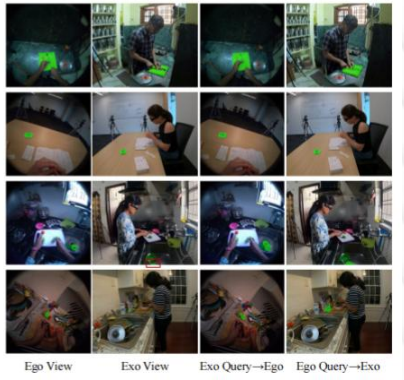
生成更高质量伪标签
进入下一次迭代

03 实验设置与结果

消融实验

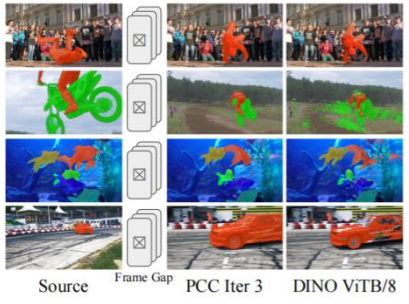
空间视角差异任务

（EgoExo4D 数据集）



时间间隔任务

（DAVIS-2017 与 LVOS 数据集）



实验一：验证不同预训练任务和解码器复杂度，是否影响模型在监督训练下的性能

Pretraining	Dec. Params.	Ego/Exo Query IoU ↑	Loc. Score ↓	CA. ↑
MAE	50M	36.9/40.2	0.043/0.078	0.57/0.51
MAE	100M	**37.7/43.7**	**0.042/0.060**	**0.60/0.56**
Grayscale	100M	31.0/34.3	0.048/0.089	0.49/0.43

解码器复杂度提升有益，灰度着色预训练不适合监督训练

实验二：验证伪标签生成时的“帧间隔长度”，是否影响模型在视频对应任务（DAVIS-2017）中的性能

Pseudolabel Temporal Gap	$\mathcal{J}\&\mathcal{F}_m$ 2 Frames	10 Frames	30 Frames
2 seconds	**83.9**	75.8	65.4
4 seconds	83.5	**76.2**	66.3
6 seconds	83.7	**76.2**	**67.7**

谢谢！