DropMAE：通过带空间注意力丢弃的掩码自编码器学习时序匹配任务的表征

Qiangqi ang Wu、Tianyu Yang、Ziquan Liu、Wei Lin、Baoyuan Wu 和 Antoni B. Chan

摘要——本文研究了掩码自编码器（MAE）视频预训练在多种基于时序匹配的下游任务中的应用，包括对象级跟踪任务（如视频目标跟踪（VOT）和视频目标分割（VOS））、自监督视觉对应关系学习、密集跟踪任务（如光流估计和长期点跟踪）以及三维点云跟踪。具体而言，我们旨在探索如何学习一种通用表征，以提升各种下游跟踪任务中的时序匹配能力。为此，我们首先发现，MAE 的一种简单扩展——在视频中随机遮蔽帧图像块并重建帧像素——过度依赖空间线索而忽略了帧重建中的时间关系，因而导致时序匹配表征不够理想。为缓解这一问题，我们提出了 DropMAE，在帧重建过程中自适应地执行空间注意力丢弃，以促进视频中的时序对应学习。通过 DropMAE，我们得出几项重要发现：1）DropMAE 是一种强大且高效的时序匹配学习器，在基于匹配的任务上取得了优于基于 ImageNet 的 MAE 的微调结果，同时具有 2× 更快的预训练速度。2）DropMAE 适用于不同的跟踪任务，包括 VOT 和 VOS 等对象级匹配任务、光流估计和任意点跟踪（TAP）等密集跟踪任务，甚至还适用于不同模态的点云数据三维跟踪。3）使用 DropMAE 预训练权重，可显著加速（例如，在 K400 上快 16.6× 倍）现有的自监督视觉对应关系学习。4）对于提升下游跟踪性能而言，预训练视频中的运动多样性比场景多样性更为重要。由于当时尚无此类模型，我们为不同的下游跟踪任务构建了基于 ViT 的跟踪器；预训练的 DropMAE 模型无需进一步修改即可直接加载到这些跟踪器中进行微调。在 6 项下游跟踪任务上的实验表明，DropMAE 作为适用于多种跟踪任务的通用预训练表征是有效的。代码和预训练模型可在以下地址获取：https://github.com/jimmy-dq/DropMAE.git。

索引词—生成式预训练、视频目标跟踪、视频目标分割、自监督对应关系学习、光流估计、长期点跟踪、深度学习

arXiv:2304.00571v3 [cs.CV] 4 Apr 2025

1 引言

变换器在多个研究领域取得了显著成功，包括自然语言处理（NLP）[1, 2],计算机视觉 [3], 和音频生成 [4, 5]。在自然语言处理中，掩码自编码已成为训练拥有数十亿参数的大规模变换器的标准方法，能够学习到稳健且具有泛化能力的表征。受自然语言处理中自监督学习成功的启发，近期的计算机视觉研究 [6, 7] 探索了适用于视觉变换器的类似策略。其中，开创性工作 MAE [6] 提出了一种掩码自编码框架，通过部分图像块重建输入图像。MAE 学习到的表征已在多种计算机视觉任务中展现出有效性，包括图像分类、目标检测和语义分割。

近期视频目标跟踪（VOT）领域的进展表明，将 MAE 预训练的 ViT 模型用作骨干网络是有效的，SimTrack [8]和 OSTrack[3]便是例证。值得注意的是，这两种跟踪器无需使用复杂的跟踪流程，便在标准基准测试中取得了最先进的性能。它们的成功可归功于 MAE 在 ImageNet [9]上获得的稳健预训练权重。此外，[8, 3]表明，对于 VOT，MAE 在 ImageNet 上的无监督 预训练优于有监督 预训练（后者使用类别标签）——这主要是因为 MAE 预训练能够有效捕捉细粒度局部结构，而这些结构对于 VOT 中的精确目标定位至关重要；相比之下，有监督预训练侧重于高层类别特定特征，这类特征具有不随外观变化而改变的特性，因此不太适用于跟踪任务。尽管[8, 3], 表现出色，但基于 ImageNet 的 MAE 预训练对于跟踪任务而言仍非最优，这是因为图像与视频之间存在固有差距：静态图像无法学习先验时间对应信息。此外，先前的方法 [10, 11, 12] 已表明，学习时间对应关系对于构建稳健且具有判别性的跟踪器至关重要。因此，我们有机会进一步针对基于匹配的下游跟踪任务专门开发 MAE 框架，例如 VOT、视频目标分割（VOS）以及密集跟踪任务（如光流估计和长时点跟踪）。

• *Qiangqiang Wu*、*Wei Lin* 和 *Antoni B. Chan*（通讯作者）均就职于香港城市大学计算机科学系。电子邮箱：*qiangqwu2-c@my.cityu.edu.hk*、*wlin38-c@my.cityu.edu.hk*、*abchan@cityu.edu.hk*。

• *Tianyu Yang* 就职于阿里巴巴达摩院，中国杭州。（电子邮箱：*tianyu-yang@outlook.com*）

• *Ziquan Liu* 就读于伦敦玛丽女王大学电子工程与计算机科学学院。（电子邮箱：*ziquan-liu.cs@gmail.com*）

• *Baoyuan Wu* 就职于香港中文大学（深圳）数据科学学院，中国广东省 *518172*。（电子邮箱：*wubaoyuan@cuhk.edu.cn*）



图 1. 适用于多种下游跟踪任务的通用 DropMAE 预训练模型，包括目标级跟踪（即 VOT 和视频目标分割）、三维点云跟踪、密集跟踪（即光流估计和长期点跟踪），以及用于无监督跟踪的自监督对应关系学习。

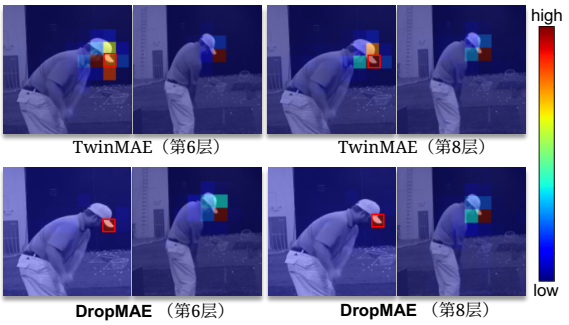

图2. TwinMAE 基线与我们提出的 DropMAE 在重建随机掩码图像块时的注意力图可视化。该图像块在左侧输入帧中以红色边界框标出。与时间线索（帧间）相比，TwinMAE 在重建时更多地利用空间线索（同一帧内）。我们提出的 DropMAE 通过有效缓解重建过程中空间线索之间的协同适应，改进了基线，使模型更加关注时间线索，从而在跟踪任务中更好地学习时间对应关系。

将 MAE 简单扩展到视频表征学习，就是对视频片段（例如帧对）中的各帧图像块进行随机掩码，并重建该片段。我们将这一简单基线称为 *TwinMAE* (TwinMAE)。如图 2和 4 所示，给定一个掩码图像块查询，TwinMAE 主要依赖同一帧内空间相邻的图像块进行重建。这意味着重建过程中过度依赖空间线索（帧内词元），并可能导致模型学到不适用于视频目标跟踪和分割等基于匹配的下游任务的次优时间表征。

为解决 TwinMAE 基线所存在的上述问题，我们提出了 DropMAE，这是一种专为基于时序匹配的视频下游任务（例如 VOT、视频目标分割、自监督对应关系学习和长期密集跟踪）中的掩码自编码器设计的预训练方法。DropMAE 会自适应地执行空间注意力丢弃，在帧重建过程中打破空间线索（帧内词元）之间的协同适应，从而促进更强的时间交互，并有助于在预训练期间学习时间对应关系。值得注意的是，我们通过 DropMAE 观察到以下几个关键发现：1）DropMAE 是一种强大且高效的时序匹配学习器，在基于匹配的任务上取得了优于基于 ImageNet 的 MAE 的微调结果，同时预训练速度提高了2× 。2）在预训练视频中，运动多样性比场景多样性对提升下游跟踪性能更为重要。3）DropMAE 适用于不同的跟踪任务，即目标级跟踪（包括 VOT 和视频目标分割）、基于像素级的稠密跟踪任务（包括光流估计和长期点跟踪）、视频中的自监督对应关系学习，甚至适用于不同模态的三维点云数据跟踪。4）将 DropMAE 用作预训练权重，可使 K400 上的自监督视觉对应关系学习提速 16.6× 。

为评估DropMAE的有效性，我们在6项下游跟踪任务上进行了实验，包括*VOT*、*VOS*、自监督视觉对应学习、密集跟踪（包括光流估计 和长期点跟踪）以及三维点云跟踪。由于某些下游任务可能缺少基于ViT的跟踪基线，我们构建了这些ViT基线以作进一步研究。对于VOT和VOS，我们发现，采用DropMAE 预训练的跟踪器在GOT-10k上取得了75.9%的平均重叠率，在LaSOT$_{ext}$上取得了52.7%的曲线下面积，在TNL2K上取得了56.9%的曲线下面积，并在DAVIS-16/17上取得了92.1%/83.0%$\mathcal{J}\&\mathcal{F}$ 分数，且无需使用复杂的在线更新或记忆机制。对于长期点跟踪，我们基于测试时优化的DropDINO跟踪器在DAVIS-上取得了65.6%的新最佳AJ分数，480 [13],优于DINO-Tracker [14]且可学习参数更少。对于三维点云跟踪，我们展示了DropMAE的表现优于3D预训练方法，同时显著优于二维MAE，展现了其在三维跟踪中的潜力。总体而言，各类下游DropMAE跟踪器所取得的出色跟踪性能，体现了我们预训练方法的优越性。

In 总之，t he mai^{n} contr ibuti^{ons} o f our wor k are:

- 据我们所知，我们是首个针对多种基于时序匹配的下游任务研究掩码自编码器视频预训练的团队，包括视频目标跟踪和视频目标分割等对象级跟踪任务、光流估计和长期点跟踪等像素级密集跟踪任务、自监督视觉对应学习以及三维点云跟踪。

- 我们探索了多种用于预训练的视频数据源，并构建了 TwinMAE 基线，以研究其在各种时序匹配任务中的有效性。

- 由于一些下游任务缺少基于 ViT 的跟踪基线，我们为视频目标分割、光流估计、长期点跟踪和自监督视觉对应学习等任务构建了 ViT 基线。这些 ViT 基线可以直接加载我们经过 DropMAE 预训练的模型进行微调，无需进一步修改。

- 我们提出了 DropMAE，它在帧重建过程中自适应地进行空间注意力丢弃，以促进视频中的有效时序对应学习。预训练的 DropMAE 模型可直接加载到我们基于 ViT 的下游跟踪器中进行微调，无需进一步修改。

- 在 13 个基准测试中的 6 项下游跟踪任务上进行的实验表明，DropMAE 是一种适用于多种跟踪任务的通用预训练表示。

我们这项工作的初步版本发表于会议论文 [15], ，当时仅研究了 VOT 和 VOS 两项下游任务。本文在会议版本的基础上，进一步研究了以下 4 项下游任务：1）通过提出用于光流估计和长期

点跟踪的 DropRAFT 和 DropDINO 跟踪器，将 DropMAE 拓展到密集跟踪；2）研究 DropMAE 在自监督视觉对应关系学习中的应用，并将其集成到现有的自监督方法中，在实现具有竞争力的无监督跟踪性能的同时，大幅提升训练速度（例如，在 K400 上快 16.6× 倍）；3）将 DropMAE 应用于三维点云跟踪，展示其时序匹配能力可泛化至不同模态；4）新增了关于密集跟踪、自监督对应关系学习和三维点云跟踪的实验。

本文其余部分组织如下。第§2节回顾相关工作。随后，我们在 §3节介绍所提出的方法。接着，§4节展示了我们提出的 DropMAE 与现有预训练方法之间的比较。在 §5节中，我们展示并讨论了 6 项下游跟踪任务的实验结果；§6 节则介绍消融研究。最后，我们在 §7节总结全文。

2 相关工作

本节回顾视频目标跟踪与分割、自监督学习、三维点云跟踪、光流估计和长期点跟踪的相关工作。

2.1 视频目标跟踪与分割。

给定测试视频第一帧中标注的边界框，视频目标跟踪（VOT）旨在准确预测后续帧中的目标边界框。类似地，对于视觉目标分割（VOS），给定第一帧中标注的二值掩码，VOS 旨在预测其余帧中的稠密目标掩码。在 VOT 发展的早期，基于相关滤波器的方法 [16, 17, 18, 19, 20]凭借出色的目标外观变化建模能力，成为主流跟踪器。随着深度学习的发展，深度孪生网络 [21] 被引入 VOT。具有代表性的工作 SiamFC [11]将模板图像和搜索图像作为输入，以实现目标定位。在 SiamFC 的基础上，研究者提出了许多改进，例如尺度回归 [12, 22],、在线模板更新[23, 24],、多层级特征融合 [25], 和骨干网络设计[3, 8, 22, 26]。对于 VOS，基于匹配的方法，如 STM[27], 、AOT [28]和 STCN [29], 在现有 VOS 基准测试中取得了良好效果。近期的改进 [30, 31] 进一步优化了在线记忆设计，使 VOS 的性能超越以往的最先进水平。SimTrack [8] 和 OSTrack [3]等近期研究表明，采用 ImageNet 上 MAE 预训练的 ViT 骨干网络 [32]有助于提升目标跟踪效果。

尽管 ViT 通过 MAE 预训练在跟踪领域取得了巨大成功，但这种基于静态 ImageNet 的预训练仍缺乏时序对应学习。此外，VOT 和 VOS 的发展表明，学习强大的时序匹配能力对视频跟踪任务至关重要。据我们所知，我们是首个研究用于跟踪 任务的掩码自动编码器自监督视频预训练的团队。我们的 DropMAE 能够为跟踪器初始化提供稳健的预训练权重，这一点已在现有跟踪器中得到验证，例如 HIPTrack[33], DiffusionTrack [34], TGTrack [35] 和 BofN [36]。

2.2 自监督学习

自监督学习在过去几十年中受到广泛关注。人们设计了许多用于预训练的前置任务，例如图像着色 [37], 拼图求解 [38], 未来帧预测 [39, 40] 和旋转预测 [41]。近年来，对比学习方法 [42, 43, 44, 45, 46, 47]已成为主流的自监督方法。同期研究 [48, 49, 50, 51, 52, 53]表明，将时间不变约束引入基于对比学习的方法，可提升视频动作识别的性能。然而，这些基于对比学习的方法对所采用的数据增强的类型和强度十分敏感，因此难以训练。受掩码语言建模 [2, 1], 的启发，研究者提出了掩码图像建模（MIM）方法，用于无监督图像学习 [6, 7]。VideoMAE [54]和 ST-MAE [55]采用立方体掩码来学习视频表征，已证明这对包括图像分类和视频动作识别在内的许多高层下游任务都很有效。SiamMAE [56] 通过掩码未来帧扩展 MAE [6] ，以执行自监督时序对应关系学习这一单一任务。然而，这些基于视频的生成式方法仍缺乏能够更有效地学习时序匹配的自适应掩码策略。此外，针对当前基于时序匹配的跟踪任务，尚无专门设计的通用预训练方法。在这项工作中，我们提出 DropMAE 来探索这一方向。我们的 DropMAE 利用自适应空间注意力丢弃，更好地促进时序学习。作为一种通用预训练模型，DropMAE 展现出强大的时序匹配自监督学习能力。

2.3 三维点云跟踪

受二维视觉目标跟踪的启发，三维单目标跟踪（SOT）旨在根据首帧中的初始三维目标边界框，在点云数据中估计三维目标边界框。P2B [57]提出采用三维区域提议网络生成三维提议框，在 [58, 59, 60] 上取得了优于此前基于形状补全的 SC3D [61]的性能，同时实现了实时速度。为利用边界框先验，BAT [62]在 P2B 的基础上引入了边界框感知特征设计。V2B [63]进一步提出在鸟瞰视图（BEV）中执行三维单目标跟踪。然而，由于匹配能力有限，这些方法在具有挑战性的场景中可能会出现性能下降。为缓解这一问题，研究者提出了基于三维 Transformer 的方法来进行目标感知特征学习。这些改进包括先进的 Transformer 架构 [64, 65, 66, 67, 68, 69],在线记忆建模 [70, 71, 72], 运动预测 [73, 74] 以及二维到三维蒸馏[75]。尽管取得了这些成果，三维单目标跟踪领域的预训练研究仍然有限。在本文中，我们证明，当三维点云训练数据有限时，从二维视频中学习得到的 DropMAE 预训练模型可以提升三维跟踪性能。

2.4 光流估计

光流估计是一项基础任务，旨在估计视频帧之间逐像素的二维运动。传统方法通过强正则化来最大化对应像素之间的视觉相似度 [76, 77, 78]。基于深度学习的方法则以端到端可训练的方式解决该任务。具体而言，FlowNets [79, 80]将该任务表述为一个密集回归问题。DCNet [81] 和 PWC-Net [82] 引入四维代价体，以显式建模视频中的像素对应关系。RAFT [83] 通过在多尺度四维代价体上应用迭代循环细化，改进了以往方法，并启发了众多后续研究 [84, 85, 86, 87, 88, 89, 90]。SEA-RAFT [91] 提出了一种更高效、精确的 RAFT

光流估计方法。FlowFormer++ [92]引入了掩码代价体自编码，利用 MAE [6] 预训练专门提升光流估计性能，其做法是对代价体进行掩码。相比之下，我们的 DropMAE 采用自适应视频帧掩码，因此能够更广泛地适用于各种下游跟踪任务。值得注意的是，在使用合成数据微调的 Sintel [93] 基准最终通道上，DropMAE 获得了更低的平均端点误差（AEPE），表明其泛化性能优于 FlowFormer++。

2.5 长期点跟踪

以往的光流方法主要关注像素位移，但在长期跟踪一致性方面表现不佳。近期研究通过利用先进的模型和数据集来解决这一问题。PIPs [94]引入了用于迭代轨迹更新的 MLP-Mixer，而 TAP-Net [95]提供了点跟踪数据集和用于位置回归的神经网络。TAPIR [96] 利用 PIPs 中的 MLP-Mixer 优化点轨迹。MFT [97] 通过分析光流不确定性和遮挡，选择可靠的光流链，以实现长期跟踪。PointOdyssey [98] 结合 PIPs++并采用时序卷积，增强了时序跟踪能力。基于Transformer的方法 [99, 100, 101, 102, 103, 104] 进一步提升了跟踪可靠性。然而，这些方法依赖于使用合成数据集进行离线训练 [95, 98],，缺乏针对特定测试视频的在线适应能力。

为适应在线测试视频，研究者提出了多种测试时优化方法。OmniMotion [105] 通过将二维像素提升至三维来优化点轨迹。DeCoMotion [106] 通过将视频分解为静态与动态成分进一步提升了这一方法。为加速OmniMotion， [107] 提出了CaDeX++，用于分解时空特征。近期，DINO-Tracker [14] 采用测试时训练，并使用预训练的DINO-ViT模型 [108], 实现了最先进的点跟踪性能。在本文中，我们展示了DropMAE凭借其鲁棒的预训练权重，能够有效地对训练数据有限的在线视频进行测试时优化。与现有方法相比，DropMAE跟踪器需要微调的参数更少，同时实现了最先进的跟踪性能。

3 方法

我们提出了一种自监督视频预训练方法，用于学习适用于基于时序匹配的下游任务的鲁棒表征，包括视频目标跟踪（VOT）、视频分割分割（VOS）、三维点云跟踪、光流估计、长期点跟踪以及自监督视觉对应关系学习。我们首先介绍一种将 MAE 简单扩展至视频时序匹配表征学习的方法，称为 TwinMAE 基线。随后，我们阐明 TwinMAE 的局限性，并提出一种促进时序对应学习的空间注意力丢弃策略，称为 DropMAE。DropMAE 和 TwinMAE 的整体流程如图 3 所示。最后，我们介绍用于微调各类下游跟踪任务的多种基于视觉 Transformer 的基线。

3.1 TwinMAE：时序掩码自动编码器基线

掩码自编码器（MAE）模型 [6] 由编码器和解码器组成。其基本思路是随机掩码图像中很大一部分（例如 75%）的图像块，然后重建图像像素。具体而言，编码器仅将可见图像块作为输入进行特征学习，随后解码器以

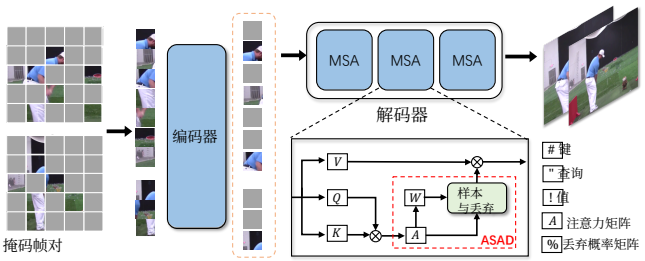

图 3. 我们的 DropMAE 示意图。所提出的自适应空间注意力丢弃（ASAD）有助于为时序匹配任务学习时间对应关系。TwinMAE 遵循相同的流程，但不使用 ASAD 模块。

可见和掩码图像块作为输入，生成图像重建结果。为了适应下游基于视频匹配的任务，一种直接的扩展方法是将 MAE 应用于拼接后的视频帧，希望从视频帧对中学习时序匹配表征，我们将其称为 TwinMAE。

值得注意的是，现有将 MAE 扩展到视频表征学习的工作 [55, 54] 主要针对视频动作识别这一下游任务设计，其中使用较长的视频片段（例如 16 帧）进行基于重建的预训练。为与我们的下游跟踪任务保持一致，我们采用目标跟踪 [3, 11] 和密集跟踪 [83, 14] , 中使用的通用训练设置，从一个视频中采样两帧作为 TwinMAE 的输入进行预训练。与现有视频预训练方法 [55, 54],相比，这种调整显著降低了计算和内存开销，因为视觉 Transformer 的复杂度呈二次增长。

图像块嵌入。首先，我们从视频中随机采样两帧，并预先设定最大帧间隔。对于每一帧，我们按照原始 ViT 的方式将其划分为互不重叠的图像块。随后，将从这两帧提取的图像块拼接起来，形成整体图像块序列。接着，我们随机掩码图像块序列中的图像块，直到达到预先设定的掩码比例。请注意，我们采用与原始 MAE 相同的掩码比例（即 75%），因为两帧中的信息冗余度应与单张图像相近。可见图像块经线性投影 [32], 嵌入，掩码图像块则使用共享的可学习掩码标记进行嵌入。所有嵌入后的图像块均加上位置嵌入 [32]。

帧身份嵌入。为了区分两帧在相同空间位置上的掩码词元，我们使用两个可学习的帧身份嵌入来标识这两帧输入帧。相应的帧身份嵌入会加到每个嵌入后的图像块上。

自动编码器与训练。遵循原始 MAE 的自动编码流程 [6], , 编码器仅将可见的嵌入图像块作为输入，而解码器则接收所有嵌入图像块，以进行掩码补丁重建。我们采用 MAE 中相同的归一化像素损失来训练整个网络。

3.2 TwinMAE基线的局限性

图中展示了我们的 TwinMAE 基线的重建可视化结果。2。我们还在图中定量比较了重建过程中的平均帧内注意力和帧间注意力。4。有趣的是，我们发现 TwinMAE 的重建严重依赖帧内图像块或空间线索，这可能导致其在基于匹配的视频任务中获得次优的时间表征。当仅使用帧内空间线索时，

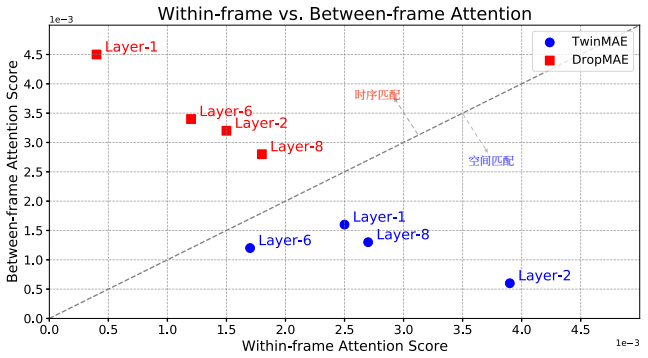

图 4. TwinMAE 和 DropMAE 在不同解码器层中获得的帧内与帧间注意力分数的平均值。注意力分数是在 20 个随机抽取的 K400 验证视频上计算的，并对所有注意力头和位置取平均。

解码器将仅利用相邻图像块中的上下文信息进行重建，因此学到的编码器表征将嵌入上下文信息。相比之下，使用帧间线索时，解码器将学习匹配不同帧中的图像块，从而恢复另一帧中对应的目标图像块。因此，利用帧间线索进行解码，会促使编码器学习支持帧间时序匹配的表征。VOT 领域的先前工作 [10, 11] 也表明，时序对应学习对于开发稳健且具有判别能力的跟踪器至关重要。由于 TwinMAE 更依赖上下文信息，因此对于下游跟踪任务而言仍非最优。

3.3 基于自适应空间注意力丢弃的 DropMAE

为了解决 §3.2 中讨论的 TwinMAE 问题，我们提出了自适应空间注意力丢弃（ASAD），以促进模型中的时序对应学习，并将该模型称为 DropMAE。对于一个查询词元，我们的基本思路是自适应地丢弃其部分帧内线索，从而帮助模型学习更可靠的时间对应关系，即帧间线索。也就是说，我们通过调整 Transformer 中计算得到的空间注意力，限制查询词元与同一帧中词元之间的交互，并鼓励其与另一帧中的词元进行更多交互。因此，为了最小化重建损失，该模型能够学习更强的时序匹配能力，这对于基于匹配的视频任务至关重要。

在介绍所提出的 ASAD 之前，我们首先回顾 ViT 中的多头自注意力 [32]。令 $\mathbf{z} \in \mathbb{R}^{N\times D}$ 表示两个拼接输入帧组成的输入序列，N 表示这两帧中的图像块总数，D 表示特征维度。标准的多头自注意力 [32] 为：

$$[\mathbf{q},\mathbf{k},\mathbf{v}] = \mathbf{z}\mathbf{U}_{qkv}, \quad \mathrm{SA}(\mathbf{z}) = \mathrm{softmax}(\frac{1}{\sqrt{D_k}}\mathbf{q}\mathbf{k}^T), \tag{1}$$

$$\mathrm{MSA}(\mathbf{z}) = [\mathrm{SA}_1(\mathbf{z}); \mathrm{SA}_2(\mathbf{z}); \cdots; \mathrm{SA}_k(\mathbf{z})]\mathbf{U}_m, \tag{2}$$

其中$\mathbf{U}_{qkv} \in \mathbb{R}^{D\times 3D_k}$，$\mathbf{U}_m \in \mathbb{R}^{k\cdot D_k\times D}$，$A = \frac{\mathbf{q}\mathbf{k}^T}{\sqrt{D_k}} \in \mathbb{R}^{N\times N}$ 和 $\mathbf{U}$。令 $\mathbf{q}\mathbf{k}$ 表示注意力矩阵。我们的 ASAD 在空间上执行-注意力丢弃以 A 移除部分帧内交互。**时间匹配概率**。我们首先需要考虑最适合应用 ASAD 的词元。直观而言，在另一帧中匹配度较高的查询词元是理想的候选对象，因为即使缺少帧内线索，它仍能



(a) 第 4 层 (b) 第 6 层

图 5. 示例帧对上的时序匹配函数 f_{tem} 的可视化。$f_{tem}(i)$ 的值越大，表示第 i 个像素与另一帧中的某个像素匹配得越好。

利用另一帧中的时间线索得到良好重建。在此，我们定义一个时序匹配函数 $f_{tem}(\cdot)$，用于衡量第 i 个查询词元的时间匹配概率：

$$f_{tem}(i) = \max_{j\in\Omega_t(i)}(\hat{A}_{i,j}), \quad \hat{A} = \mathrm{softmax}_{\mathrm{row}}(A), \tag{3}$$

其中，softmax 函数按 A,$f_{tem}(i) \in [0,1]$的每一行计算；$\Omega_t(i)$ 表示时序索引集合，包含另一帧中所有词元的索引，且这些词元均为第 i 个查询词元。(i) 的值越大，表示第 i 个查询词元与另一帧匹配良好的概率越高，因此它是 ASAD 的理想候选。图 5 展示了 $f_{tem}(\cdot)$ 的可视化。第 5 图。**总体丢弃概率的测量**。位置(i,j) 处的总体空间注意力丢弃概率由时序匹配概率和归一化空间重要性共同计算：

$$W_{i,j} = f_{tem}(i)\frac{\hat{A}_{i,j}}{\sum_{j'\in\Omega_s(i)}\hat{A}_{i,j'}}, \quad j \in \Omega_s(i), \tag{4}$$

其中 $\Omega_s(i)$ 是空间索引集，包含与第 i 个查询处于同一帧中的所有其他词元索引（即不包括查询自身的索引）。当 $W_{i,j}$ 较大时，第 i 个查询词元具有良好的帧间匹配，同时，第 j 个帧内词元对第 i 个查询也很重要。在这种情况下，丢弃 A 中的帧内注意力元素 (i,j) 有助于模型利用帧间（时间匹配）词元进行词元学习或重建。最后，我们将自注意力和时间自注意力的丢弃概率设为 0，即 $W_{i,i} = W_{i,i+N/2} = 0$。

请注意，总共有 $N(N/2-1)$ 个空间注意力元素（即不包括自注意力元素），且只有这些空间注意力元素会参与丢弃。给定预设的丢弃率 P，我们从 A 中全局丢弃总计 $N_d = PN(N/2-1)$ 个注意力元素。

Dropout 的采样。我们根据丢弃概率矩阵 N_d，从多项分布中抽取 W 个元素。然后，我们将 A 中对应索引的元素设为 $-\infty$。在（1）中应用 softmax 函数后，相应的空间注意力权重会被移除。其余操作与 ViT 中使用的原始多头自注意力机制相同。类似 PyTorch 的伪代码见算法1。

自动编码器与训练。与 TwinMAE 相比，我们的 ASAD 方法由于在图形处理器上采用了高效的矩阵运算，几乎不会增加额外的时间成本。在预训练阶段，我们将 ASAD 应用于解码器的每一层，以学习支持时序匹配的编码器表示。如图 4 所示，我们采用 ASAD 的 DropMAE 利用了更多的 帧间注意力进行重建，从而学会执行精确的时序匹配，以恢复帧间图像块，进而实现更好的时序对应关系学习。

算法 1： ASAD 伪代码，PyTorch 风格

```
# 输入：注意力矩阵 A，序列长度
 N，丢弃数量 N_d
W  = torch.zeros_like(A)# N×N
A = A.detach().softmax(dim=-1))  # N×N

# 获取 A 每一行中的时间注意力
A_tem = 时间_index(A)            # N×N//2
f_时 = A_时。最大值(维度=-1).值 # N-by-1

# 获取 A 每一行中的空间注意力
A_空间 = 空间_索引(A)            # N×N//2
# 避免自注意力丢弃
A_空间[0:N//2,    0:N//2]。填充_对角线_(0)
A_spa[N//2:,      0:N//2].填充_对角线_(0)
A_spa=A_spa/A_spa.求和(维度=-1，保持维度=True)

# 计算总体丢弃概率
f_全部 = f_tem * A  _spa#N-by-N//2

# 放回概率矩阵 W
W[0:N//2 ，0:N//2]  = f_all[0:N//2   ，0:N//2]
W[N//2:,   N//2:] = f_全部[N//2:,      0:N//2]
基于 W 采样 N_d 个元素

索引=torch.multinomial(W.view(1,-1),N_d)返回 索引
```

在下一节中，我们将介绍基于预训练 ViT 模型的下游任务微调。

3.4 下游时序匹配任务

获得预训练的 DropMAE 模型后，我们会在下游时序匹配任务上微调训练良好的编码器（即 ViT 模型）。为展示 DropMAE 学习到的表征的通用性，我们在此考虑 6 项下游任务：视频目标跟踪、视频目标分割、三维点云跟踪、光流估计、长期点跟踪和自监督视觉对应学习。

3.4.1 视频目标跟踪

近期，在 ImageNet 上预训练的 MAE视觉Transformer模型被应用于 VOT，并取得了令人印象深刻的结果。我们使用具有代表性的跟踪器 OSTrack [3] 作为微调的基线跟踪器。在 OSTrack 中，裁剪后的模板和搜索图像首先被序列化为序列，并连接在一起。随后，将位置嵌入添加到整个序列中，并输入视觉Transformer骨干网络，以进行联合特征提取和交互。最后，将更新后的搜索特征输入预测头，以预测目标边界框。

在微调阶段，我们使用预训练的 DropMAE编码器权重初始化 OSTrack 中使用的视觉Transformer骨干网络。同时，为了与预训练阶段保持一致，分别向模板嵌入和搜索嵌入中添加两个帧身份嵌入。我们采用原始 OSTrack 的相同训练损失。我们将这一基于 DropMAE 的跟踪器称为 DropTrack。

3.4.2 视频对象分割

对于视频目标分割（VOS），目前还没有基于 ViT 的方法。因此，我们构建了一个采用视觉Transformer骨干网络的简单 VOS 基线，即 DropSeg，以弥补这一空缺。DropSeg 的整体流程如图6所示。

输入序列化。给定一个带有二值掩码的模板帧，视频目标分割旨在对视频中每一帧的感兴趣对象进行分割。

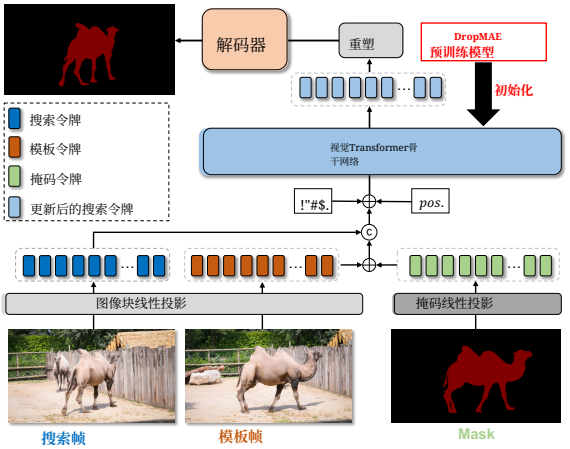

图6. 所提出的用于 VOS 的 DropSeg 整体流程。*iden* 和*pos* 分别表示帧身份嵌入和位置嵌入。

与预训练阶段类似，二值掩码图、模板帧和搜索帧首先被转换为图块序列，然后进行线性投影并加入位置嵌入。向模板嵌入和搜索嵌入中分别添加两个帧身份嵌入，并将掩码嵌入添加至模板嵌入中以进行掩码编码。

联合特征提取与交互。模板嵌入与搜索嵌入拼接后，输入视觉Transformer骨干网络，以进行联合特征提取和匹配。我们使用从 ViT 最后一层提取的更新后搜索特征进行掩码预测。

掩码预测。现有的视频目标分割方法 [27, 29, 30, 28]采用多分辨率特征进行掩码预测。然而，更新后的搜索特征只有单一分辨率。我们遵循 [109]，通过两个反卷积模块将搜索特征上采样至 2× 和 4× 尺寸。最后，我们使用 [29, 27]中相同的解码器进行掩码预测。

训练损失。我们使用常用的交叉熵损失 [29,27] 来训练整个网络架构。

在线推理。 在线推理期间，我们使用带有掩码标注的第一帧作为记忆帧，以便在搜索帧中在线匹配目标。

3.4.3 三维点云跟踪

给定目标在第一帧中的三维边界框，三维单目标点云跟踪旨在估计目标在后续帧中的边界框。这项任务成功的关键在于，在目标点云与搜索点云数据之间进行稳健匹配。我们在此研究 DropMAE 在三维匹配学习中的有效性。具体而言，我们主要遵循 SiamDisst [75], 的方法，该方法采用 ViT 作为骨干网络，用于联合特征提取和匹配，并使用三维边界框头进行边界框回归。我们使用 DropMAE 预训练权重初始化三维跟踪器 SaimDisst 中的视觉Transformer骨干网络，以实现更稳健的模板匹配。有趣的是，我们发现 DropMAE 预训练模型对于三维点跟踪同样有效，尽管该模型仅在二维视频上进行预训练。我们希望这项研究能启发更多关于将二维预训练模型迁移到三维跟踪任务的研究。

3.4.4 光流估计

光流专注于估计连续帧之间的短期稠密运动。RAFT [83] 是一种典型的光流

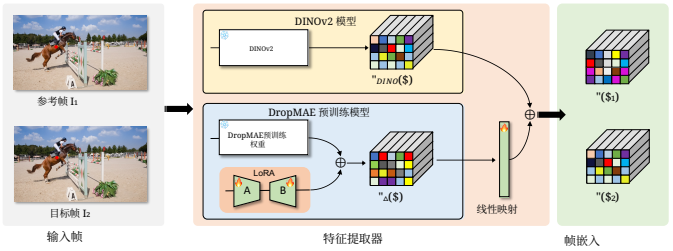

图 7. 所提出的 DropDINO 整体流程：采用参数高效的 LoRA 训练，并结合我们的 DropMAE 预训练模型进行长期点跟踪。

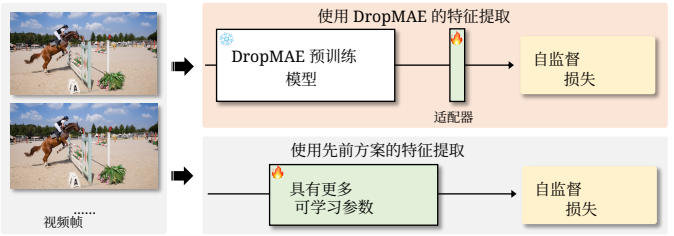

图 8. 使用我们的 DropMAE 预训练模型进行参数高效的自监督表征学习的整体流程。

方法，通过迭代细化预测逐像素的二维向量场。具体而言，RAFT [83] 分三个主要步骤预测相邻两幅 RGB帧之间的稠密二维光流场：(1) 使用特征编码器和上下文编码器从输入图像中提取低分辨率特征图；2) 通过计算两幅输入帧中特征图对之间的视觉相似度，构建完整的相关体积图；3) 使用循环神经网络单元迭代细化光流预测。

本文采用 RAFT 作为基线，结合 DropMAE 预训练模型，以实现更好的时序匹配。在 RAFT 中，特征提取和匹配是分开进行的。RAFT 使用基于 ResNet 的特征编码器 [110] 首先提取输入帧的特征，然后采用人工设计的匹配层（即通过矩阵乘法实现）计算视觉相似度图。然而，这种分离式方案无法提取动态目标感知特征，因而无法更好地计算视觉相似度。为解决这一问题，我们可以为 RAFT 配备 ViT 特征提取器，以提取目标感知特征。然而，在初步实验中，直接将 ViT 与一些现成的预训练权重（例如随机初始化权重、监督式 ImageNet 权重和 MAE [6]权重）结合用于 RAFT，会导致性能显著下降（见表9），这主要是因为这些预训练权重缺乏丰富的时间先验。

为缓解这一问题，我们使用 DropMAE 预训练模型作为特征编码器，使 RAFT 能够有效利用基于 ViT 的特征提取器。新的跟踪器（记作 DropRAFT）使 RAFT 能够对两个输入帧进行联合特征提取与匹配，从而生成更准确的相关体积，并显著提升最终性能（见表 9）。整个过程可表示为：

$$I_1', I_2' = E_D(I_1, I_2), \tag{5}$$

$$C_k = \text{AvgPool}(I_1' \cdot {I_2'}^T, 2^k) \in \mathbb{R}^{H \times W \times \frac{H}{2^k} \times \frac{W}{2^k}}, \tag{6}$$

其中，$E_D(\cdot,\cdot)$ 是我们的预训练 DropMAE 编码器，它将帧 I_1 和 I_2 的拼接图像块作为输入，进行联合特征提取与交互，得到 $I' \in \mathbb{R}^{H \times W \times D}$。$C_k$ 是第 k个相关体积，通过使用核大小为的平均池化得到

$\{2^k\}_{k=0}^3$在 $I_1' \cdot {I_2'}^T \in \mathbb{R}^{H \times W \times H \times W}$ 的最后两个维度上。获得 $\{C_k\}_{k=0}^3$后，我们使用 RAFT 中相同的迭代细化方法来获得光流预测。在这项工作中，我们展示了 DropMAE 得到的更准确的相关体积 $I_1' \cdot {I_2'}^T$ 能够显著提高光流估计的精度。

3.4.5 长期点跟踪

长期点跟踪旨在匹配视频中相距较远的帧之间的对应点。TAP-Vid [95] 提出了将这一问题表述为任意点跟踪（TAP），并使用基于合成数据集的卷积神经网络基线来解决。近期，Dino- Tracker [14] 将在线学习的残差卷积神经网络模型与预训练的 DINO-ViT 模型相结合 [108]，实现了最先进的长期点跟踪性能。

重新介绍 Dino-Tracker。给定初始视频帧中的一个查询点，任意点跟踪旨在跟踪后续帧中的查询点，并准确估计其长期轨迹和遮挡状态。为适应特定的在线跟踪视频，Dino-Tracker 在线优化一个残差卷积神经网络模型，并结合预训练的 DINO-ViT 模型使用。Dino-Tracker 的核心思想是通过残差卷积神经网络模型预测相对于预训练 DINO-ViT 模型的残差。残差表征旨在有效捕捉时序对应关系，与 DINO 特征中的语义表征形成互补。特征组合可表示为：

$$\phi(I) = \phi_{DINO}(I) + \phi_\Delta(I), \tag{7}$$

其中 $\phi_{DINO}(I)$ 和 $\phi_\Delta(I)$ 分别是 DINO 特征和残差特征。

局限性。DINO-Tracker 将 $\phi_\Delta(\cdot)$ 实现为卷积神经网络模型（即 ResNet），并声称该卷积神经网络模型能够有效利用其归纳偏置，将不同帧中相似的 RGB 图像块编码为相似的特征表征。此外，对于每个在线测试视频，$\phi_\Delta(\cdot)$均采用零初始化。然而，基于卷积神经网络的 $\phi_\Delta(\cdot)$ 存在以下局限：1）卷积神经网络的表达能力有限，可能导致对相似点或干扰点的匹配不准确；2）零初始化的 $\phi_\Delta(\cdot)$ 可能不适合在训练数据有限的情况下进行在线测试视频适配。为解决上述问题，我们提出使用 DropMAE 预训练模型作为 $\phi_\Delta(\cdot)$ 进行在线时序匹配学习。我们将这一新跟踪器称为 DropDINO。我们发现：1）在在线适应的低数据量场景下，DropMAE 是一种强大的时间学习器，能够很好地补充 Dino-Tracker 中使用的 DINO 特征；2）微调的参数比基于卷积神经网络的 $\phi_\Delta(\cdot)$ 更少，却能取得更好的跟踪性能，证明了 DropMAE 的有效性。

参数高效的 LoRA 训练。由于 DropMAE 采用 ViT-base 模型，其模型复杂度高于原始 Dino-Tracker 中基于卷积神经网络 $\phi_\Delta(\cdot)$ 的模型，因此我们采用 LoRA 训练 [111] 以实现参数高效训练，如图7所示。需要注意的是，[14] 曾尝试将 $\phi_\Delta(\cdot)$ 实现为采用 LoRA 微调的 DINOv2 ViT 模型。然而，性能显著下降，主要原因是 DINOv2 ViT 中丰富的语义特征并不适合时间微调。相比之下，得益于良好学习到的时间先验，DropMAE 通过采用参数高效的 LoRA 训练实现了最先进性能。

表 1 在 GOT-10k 上针对下游 VOT 和 VOS 任务的预训练方法比较 [112]以及 DAVIS-17 [13]。所有方法均采用 ViT-B/16 模型 [32]，使用 224×224 输入图像进行预训练。预训练时间在 64 个 NVIDIA V100 GPU 上测量。最佳的两项结果分别以**红色**和**蓝色**标示。

方法	预训练数据	轮数	预训练时间（小时）	GOT-10k（VOT）			DAVIS-17（视频目标分割）		
				AO	$SR_{0.5}$	$SR_{0.75}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
无预训练	-	-	-	62.7	72.8	53.7	69.5	66.9	72.2
监督式 IN1k [113]	IN1K	300	-	69.7	79.0	65.6	78.0	74.8	81.1
监督式 IN21k [114]	IN21K	80	-	70.2	80.7	65.4	78.5	75.4	81.7
CLIP [115]	IN1K	32	-	67.4	76.8	60.0	73.6	70.5	76.7
MOCO-v3 [116]	IN1K	300	-	70.1	80.1	65.3	78.4	75.4	81.5
BeiT [117]	IN1K	800	103.1	67.4	76.8	60.0	76.1	72.7	79.4
MAE [6]	IN1K	1600	84	73.7	83.2	**70.8**	81.7	78.5	84.9
VideoMAE [54]	K400	1600	123.4	61.6	72.7	48.4	-	-	-
TwinMAE	K400	400	20.7	72.2	83.2	65.9	79.3	76.4	82.3
TwinMAE	K400	800	41.3	72.9	83.6	68.5	80.7	77.9	83.6
TwinMAE	K400	1600	82.7	74.2	84.9	69.4	81.2	78.1	84.2
DropMAE	K400	400	21.1	73.2	83.9	67.5	81.3	78.5	84.0
DropMAE	K400	800	42.2	74.8	85.4	70.5	82.7	**79.7**	85.6
DropMAE	K400	1600	84.4	**75.8**	**86.4**	**72.0**	**83.1**	**80.2**	**86.0**
DropMAE	K700	800	92.4	**75.9**	**86.8**	**72.0**	**83.0**	**80.2**	**85.7**

3.4.6 自监督视觉对应关系学习

生成式 MAE 预训练及其后续发展主要侧重于学习用于下游任务微调的表征。在前面的几个小节中，我们考察了 DropMAE 预训练在基于微调的下游时序匹配任务中的有效性。为求完整，我们进一步评估了未经监督式微调的原始 DropMAE 表征在无监督跟踪任务中的表现。这里我们还展示了 DropMAE 可作为现有自监督学习方法 DUL 的有效预训练权重 [53]。值得注意的是，将 DropMAE 用作 [53] 的预训练权重，可显著加快视频中的自监督视觉对应关系学习。

通过 DropMAE 计算亲和矩阵。亲和矩阵表示视频中连续帧所提取视觉特征之间的相似度。目前，自监督学习方法利用亲和矩阵中的概率来寻找时间对应关系。具体而言，给定一对连续视频帧 I_t 和 I_{t+1}，我们首先使用 DropMAE 预训练模型提取帧特征，得到 I'_t 和 $I'_{t+1} \in \mathbb{R}^{H\times W\times D}$。对于第 i 个查询令牌特征 $\mathbf{q}_t^i$，其与第 t 帧中第 j 个特征之间的亲和度 $\hat{}k_t^{t+1}(i,\ j)$ 可以计算如下：

$$\hat{k}_t^{t+1}(i,j) = \text{Softmax}(I'_t, I'_{t+1}, \tau)_{i,j} = \frac{\exp(\mathbf{q}_t^i \odot \mathbf{q}_{t+1}^j/\tau)}{\sum_{s=1}^{N}\exp(\mathbf{q}_t^i \odot \mathbf{q}_{t+1}^s/\tau)}, \tag{8}$$

其中，$\hat{}K_t^{t+1} = [\hat{}k_{i,\ j}]_{N\times N}$ 是帧亲和矩阵，N 是帧中的空间特征数量，$\odot$ 表示内积，τ 是温度超参数。

通过 $\hat{}K_t^{t+1}$ 进行标签传播。$\hat{}K_t^{t+1}$中的每一行表示最后一帧中的一个查询特征与当前帧中各个词元之间的相关性。为公平比较学习到的表征，我们遵循 [118, 119]，将前一帧中不同类型的标签传播到当前帧。具体而言，给定第 t帧中的掩码标签 $M \times \mathbb{R}^{N\times 1}$，可通过 $K_t^{t+1}M$获得第$(t+1)$帧中的预测掩码。掩码传播遵循递归过程，其中当前帧的输出掩码作为后续帧的输入。如 [118, 119], 所述，我们仅保留每一行的前10个值，并将其他值置零。对于边界框和姿态标签传播，遵循 [120] 和 [119], ，我们使用高斯置信图和SiamFC进行无监督姿态与目标跟踪。

使用 DropMAE 进行自监督学习。对比学习方法 [53] 已在自监督对应关系学习中展现出优异性能。有趣的是，我们发现，DropMAE 预训练模型可以有效地与现有对比学习框架结合，从而实现高效的自监督对应关系学习。如图8所示，我们采用 DUL [53] 作为基线，并使用 DropMAE 预训练模型替换其特征提取器。此外，我们在 DropMAE 特征之上添加了一个轻量级适配器，并使用[53] 中的自训练和跨视图一致性损失对其进行训练，同时冻结 DropMAE骨干网络。值得注意的是，我们的 DropMAE 变体在 K400 上实现了 16。6× 倍的自监督学习加速，同时学习到比基线更有效的表征 [53]，且参数更少，从而实现高效的自监督对应关系学习。

4 预训练实验

本节将在 VOT 和 VOS 任务上开展实验，将我们的 DropMAE 与其他预训练方法进行比较。

4.1 实现细节

预训练。在预训练阶段，我们探索了多种大规模视频数据源来预训练DropMAE模型，包括Kinetics-400 [144] (K400)、Kinetics-600 [145](K600)、Kinetics-700 [146] (K700)、Moments in Time [147] (MiT)和WebVid-2M [148]。使用不同预训练数据集的详细性能对比见第6节的消融研究。按照原始MAE[6]所采用的训练设置，我们使用标准ViT-B/16 [32] 作为预训练的骨干网络。对于丢弃率 P，我们根据图 $P = 0.1$ 13中的消融研究进行设置。13。预训练使用64块NVIDIA V100 GPU进行。如表1所示，在K400 [144],上进行1600轮预训练约需84小时；使用64块NVIDIA A100 GPU可进一步缩短至58小时。

VOT。我们使用 LaSOT [123],COCO [149],TrackingNet [150] 和 GOT-10k [112]的训练集划分来训练基于 DropMAE 的跟踪器（记为 DropTrack）。对于 GOT-10k 评估，我们遵循单次评估设置，仅使用 GOT-10k 的训练集划分对模型进行微调。我们采用基础学习率

表 2 与四个大规模数据集上的最先进的视觉目标跟踪方法进行比较。前两项最佳结果分别以**红色**和**蓝色**标出。对于 GOT-10k 评测，所有方法均遵循单次训练协议，仅使用 GOT-10k 的训练集进行训练。我们的 DropTrack-B384 无需采用复杂的时序更新（TU），即可达到最先进性能。

方法	TU	来源	GOT-10k [112]			TNL2K [121]		LaSOT$_{ext}$ [122]			LaSOT [123]		
			AO	$SR_{0.5}$	$SR_{0.75}$	AUC	P	AUC	精度$_{Norm}$	P	AUC	精度$_{Norm}$	P
MDNet [124]	✓	CVPR16	29.9	30.3	9.9	-	-	27.9	34.9	31.8	39.7	46.0	37.3
ECO [125]	✓	2017年国际计算机视觉大会	31.6	30.9	11.1	32.6	31.7	22.0	25.2	24.0	32.4	33.8	30.1
DiMP [126]	✓	2019年国际计算机视觉大会	61.1	71.7	49.2	44.7	43.4	39.2	47.6	45.1	56.9	65.0	56.7
SiamR-CNN [127]	✓	[illegible]	64.9	72.8	59.7	52.3	52.8	-	-	-	64.8	72.2	-
LTMU [128]	✓	[illegible]	-	-	-	48.5	47.3	41.4	49.9	47.3	57.2	-	57.2
Ocean [129]	✓	ECCV20	61.1	72.1	47.3	38.4	37.7	-	-	-	56.0	65.1	56.6
TrDiMP [130]	✓	CVPR21	67.1	77.7	58.3	-	-	-	-	-	63.9	-	61.4
AutoMatch [131]	✓	ICCV21	65.2	76.6	54.3	47.2	43.5	37.6	-	43.0	58.3	-	59.9
STARK [132]	✓	ICCV21	68.8	78.1	64.1	-	-	-	-	-	67.1	77.0	-
KeepTrack [133]	✓	ICCV21	-	-	-	-	-	48.2	-	-	67.1	77.2	70.2
MixFormer-L [134]	✓	CVPR22	70.7	80.0	67.8	-	-	-	-	-	70.1	79.9	76.3
UAST [135]	✓	ICML22	63.5	74.1	51.4	-	-	-	-	-	57.1	-	58.7
AiATrack [136]	✓	ECCV22	69.6	80.0	63.2	-	-	46.8	54.4	54.2	49.6	56.9	49.1
CIA50 [137]	✓	ECCV22	67.9	79.0	60.3	50.9	**57.6**	-	-	-	66.2	-	69.6
MixFormer-22k [134]	✓	CVPR22	70.7	80.0	67.8	-	-	-	-	-	69.2	78.7	74.7
SeqTrack-B384 [138]	✓	CVPR 2023	74.5	84.3	71.4	56.4	-	50.5	61.6	57.5	71.5	81.1	77.8
ARTrack-B384 [139]	✓	CVPR 2023	**75.5**	84.3	**74.3**	-	-	**51.9**	**62.0**	**58.5**	**72.6**	**81.7**	**79.1**
SiamFC [11]		ECCVW16	34.8	35.3	9.8	29.5	28.6	23.0	31.1	26.9	33.6	42.0	33.9
SiamPRN++ [22]		CVPR19	51.7	61.6	32.5	41.3	41.2	34.0	41.6	39.6	49.6	56.9	49.1
TransT [140]		CVPR21	67.1	76.8	60.9	50.7	51.7	-	-	-	64.9	73.8	69.0
SBT [141]		CVPR22	70.4	80.8	64.7	-	-	-	-	-	66.7	-	71.1
SwinTrack-384 [142]		NeurIPS22	72.4	80.5	67.8	55.9	57.1	49.1	-	55.6	71.3	-	76.5
SimTrack-L [8]		ECCV22	69.8	78.8	66.0	55.6	55.7	-	-	-	70.5	79.7	-
OSTrack-384 [3]		ECCV22	73.7	83.2	70.8	55.9	56.7	50.5	61.3	57.6	71.1	81.1	77.6
OneTracker [143]		[illegible]	-	-	-	58.0	59.1	-	-	-	70.9	79.9	76.5
DiffusionTrack-B256（2） [34]		[illegible]	75.2	**85.9**	**72.0**	**56.5**	57.3	-	-	-	70.7	80.0	77.3
DropTrack-B384		**Ours**	**75.9**	**86.8**	**72.0**	**56.9**	**57.9**	**52.7**	**63.9**	**60.2**	**71.8**	**81.8**	**78.1**

表 3 与最先进的视觉目标跟踪方法在 OTB100 [152], ITB [153] 和 TrackingNet [150]上的比较。前两项最佳结果分别以**红色** 和 **蓝色**显示。

方法	OTB100 AUC	ITB AUC	TrackingNet 曲线下面积	TrackingNet 精度$_{Norm}$
SiamFC [11]	58.3	44.1	57.1	66.3
Ocean [129]	68.4	47.7	-	-
ATOM [154]	68.3	47.2	70.3	77.1
DiMP [126]	53.7	339	74.0	80.1
TransT [140]	**69.5**	54.7	81.4	86.7
STARK [132]	68.1	57.6	82.0	86.9
OSTrack [3]	-	**64.8**	**83.9**	**88.5**
DropTrack	**69.6**	**65.0**	**84.1**	**88.9**

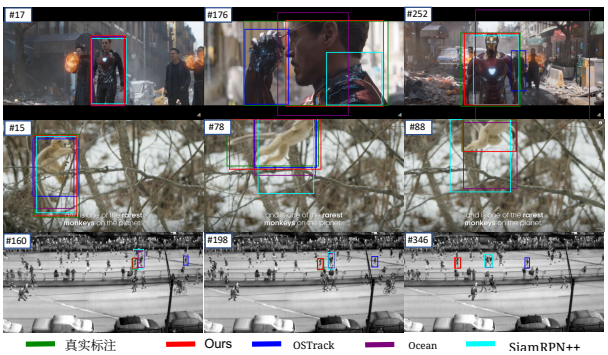

图 9. 我们的 DropTrack 与几种对比方法（包括 OSTrack [3], Ocean [129] 和 SiamRPN++ [22]）在 VOT 上的定性结果。三个视频序列取自 TNL2K [121]。每帧的编号显示在左上角 。

2.5e-4，同时将其他参数保持与 OSTrack 相同[3]。我们的 DropTrack 推理速度与基线 OSTrack 相同，在单个图形处理器上测得为 58.1 帧/秒。**视频目标分割**。 我们使用 YouTube-VOS [151] 和 Davis [13] 数据集进行微调，遵循标准做法 [27, 29]。我们使用 Adam优化器进行优化，学习率为 2e-5。模型训练 210,000 次迭代，并在 125,000 次迭代时衰减学习率。微调在 8 个 A100 GPU 上进行，整个训练过程约需 16 小时。

4.2 与预训练方法的比较

在表1中，我们将 DropMAE 与现有预训练方法在 VOT 和 VOS 下游任务上的表现进行了比较。DropMAE 和 TwinMAE 使用视频（K400、K700）进行预训练，而 MAE 和其他方法则在 ImageNet 1k 或 21k（IN1K、IN21K）上进行预训练。第 3.4 节所示的 VOT 和 VOS 基线使用现有预训练方法提供的官方预训练 VIT-B/16 模型（见表3.4 1）进行微调。经过 800 轮训练后，TwinMAE 在 VOT 上的表现优于 MAE，但在 VOS 上的结果较差。主要有两个原因：1）TwinMAE 学习时序匹配的能力还不够强；2）K400 中的目标类别数量有限，而 DAVIS-17 中的目标类别包含在 ImageNet 中。因此，MAE 在 VOS 上具有良好的泛化能力。我们的 DropMAE 是一种更强的时序匹配学习器，在使用 K400 数据集训练 800 轮（即 42.2 小时）后，在 VOT 和 VOS 任务上的表现均优于 MAE。这表明我们的 DropMAE 比 MAE 快 2× 。

5 下游任务实验

本节将我们针对六项下游任务微调后的模型，与其在各种基准测试上的最先进方法进行比较。对于 VOT 和 VOS 微调，我们使用在 K700 上训练 800 个 epoch 的 DropMAE 作为预训练模型。对于其他下游任务，为了与其他基于视频的自监督方法进行公平比较，我们使用在 K400 上训练 1600 个 epoch 的 DropMAE 作为预训练模型。

表 4 与验证集上最先进的视频目标分割方法的比较：DAVIS-2016 [155] 和 DAVIS-2017 [13]。OL、M 和 S 分别表示**O**nline **L**earning，即在线学习；使用**M**emory mechanism，即记忆机制；以及使用**S**ynthetic videos，即合成视频进行预训练。

方法	在线学习	记忆机制	合成视频	DAVIS-2016 [155]			DAVIS-2017 [13]		
				$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
RANet [156]			✓	85.5	85.5	85.4	65.7	63.2	68.2
STM [27]		✓	✓	89.3	**88.7**	89.9	81.8	79.2	84.3
FRTM [157]	✓	✓		83.5	83.6	83.4	76.7	73.9	79.6
TVOS [158]		✓		-	-	-	72.3	69.9	74.7
LWL [159]	✓	✓		-	-	-	81.6	79.1	84.1
CFBI [160]		✓		89.4	88.3	**90.5**	81.9	79.1	84.6
UniTrack [120]		✓		-	-	-	-	58.4	-
STCN⁻ [29]		✓		-	-	-	**82.5**	79.3	**85.7**
SSTVOS[161]		✓		-	-	-	**82.5**	**79.9**	85.1
SWEM⁻ [31]		✓		**89.5**	-	-	81.9	-	-
RTS [162]	✓	✓		-	-	-	80.2	77.9	82.6
OSMN [163]				73.5	74.0	72.9	54.8	52.5	57.1
FAVOS [164]				81.0	82.4	79.5	58.2	54.6	61.8
VideoMatch [165]				-	81.0	-	56.5	-	-
SiamMask [166]				69.8	71.7	67.8	56.4	54.3	58.5
D3S [167]				74.0	75.4	72.6	60.8	57.8	63.8
Siam R-CNN [167]				-	-	-	70.6	66.1	75.0
Unicorn [168]				87.4	86.5	88.2	69.2	65.2	73.2
OneTracker [143]				88.9	88.1	89.7	82.5	79.4	85.6
DropSeg				**92.1**	**90.9**	**93.3**	**83.0**	**80.2**	**85.7**



图 10. 我们的一次性方法 DropSeg 在 DAVIS-17 [13]，的四个序列上的定性 VOS 结果，分别为 *bike-packing*、*bmx-trees*、*india* 和 *soapbox*。每帧的帧编号显示在其 左上角，第一帧中给出了真实掩码。

5.1 实现细节

三维点云跟踪。 遵循 [75]，，我们使用 ViT-Base 模型的前 6 层作为骨干网络，用于联合三维点特征提取与交互。骨干网络使用我们的 DropMAE 预训练模型初始化，随后在 KITTI 数据集的厢式货车类别上进行微调 [58]，并采用与 [75]相同的训练超参数。

自监督对应关系学习。 我们将 DUL 中的骨干网络替换为 [53] 我们的 DropMAE 预训练模型，并添加了一个可学习适配器，该适配器实现为轻量级残差块（包含 5.4M 个参数，详见表 7），用于表示学习。除了由于采用了具有鲁棒性的预训练权重而仅使用一个训练周期外，我们采用了与 DUL 相同的训练设置。

光流估计。我们使用 DropMAE 预训练模型作为 RAFT 中的特征编码器 [83]。DropMAE 提取的特征会进一步上采样至 2×，以匹配 RAFT 原始特征的空间分辨率。在训练时，我们首先在 FlyingChairs 上预训练 DropRAFT [79]，，然后在 FlyingThings3D 上训练 [169]，，遵循 [83] 中相同的训练步骤和训练超参数。

表 5 采用有限训练样本时 KITTI-Van 上的三维跟踪结果。为进行比较，纳入了配备不同类型骨干网络和预训练模型的方法。

方法	骨干网络	预训练	Type	成功率	精确率
P2B [57]	PointNet++	-	-	40.8	48.4
BAT [62]	PointNet++	-	-	52.4	67.0
DMT [74]	PointNet++	-	-	53.3	65.6
M2Track [73]	PointNet	-	-	53.8	70.7
STNet [67]	Transf.	-	-	58.0	70.6
MBPTrack [71]	Transf.	-	-	61.3	72.7
SiamDisst [75]	ViT	Recon [170]	3D	62.9	73.6
SiamDisst [75]	ViT	Point-MAE [171]	3D	**63.5**	**75.0**
SiamDisst [75]	ViT	MAE [6]	2D	60.5	69.9
SiamDisst [75]	ViT	**DropMAE**	2D	**61.9**	**74.0**

长期点跟踪。我们使用 DropMAE 预训练模型作为 [14] 中用于在线像素级对应学习的特征提取器（即 Delta-DINO），并添加一个 1 × 1卷积层，以解决 DropMAE 与原始 Delta-DINO 之间的通道维度不匹配。为进行公平比较，我们对 DropMAE 使用与 DINO-Tracker 相同的训练超参数和 CNN-refiner。对于 DropMAE，我们采用 LoRA 训练进行参数高效微调，详情见表 10。按照 [14],，我们将 LoRA 的 alpha 设为 =0.5，LoRA 的 dropout 设为=0.1，秩设为=8，以进行 LoRA 训练。

5.2 视频目标跟踪

为展示所提出的 DropMAE 在 VOT 中的有效性，我们在 7 个具有挑战性的跟踪基准测试上，将 DropTrack 与最先进的跟踪器进行比较。

GOT-10k。 GOT-10k [112] 是一个具有挑战性的数据集，采用单次评估协议，要求跟踪器在其训练集划分上进行训练，且测试对象类别与训练集划分中的对象没有重叠。如表 2所示，我们的 DropTrack 在该数据集上取得了最先进的结果，在平均重叠率和 SR 方面分别超过 OSTrack 2.2% 和 3.6%$_{0.5}$。这表明，预训练中的时序对应学习有益于下游跟踪任务。尽管预训练数据与测试数据之间存在领域差距（即 GOT-10k 中很大一部分测试对象是动物、车辆和物体部件，而 K700 仅包含以人为中心的动作视频），DropMAE 学到的时序匹配能力仍可迁移到下游跟踪任务，从而提升跟踪性能。

LaSOT。LaSOT 由 280 个长测试序列组成，我们的结果展示在表2中。我们的 DropTrack 在该数据集上以 71.8% 的曲线下面积、81.8% 的精度$_{Norm}$ 和 78.1% 的精度创下新纪录，展现了 DropTrack 在稳健的长期视觉跟踪中的巨大潜力。

LaSOT$_{ext}$。 LaSOT$_{ext}$ 是 LaSOT 的扩展版本，包含更具挑战性的视频序列用于测试。与 GOT-10k 类似，测试集划分与训练集划分之间存在较大差异，并使用包含新类别物体（即 ImageNet 中未出现的物体）的序列进行评估。我们的 DropMAE 大幅超越了其他跟踪器。具体而言，在不采用复杂记忆模块设计的情况下，DropTrack 在曲线下面积、精度 [139]和精度指标上分别超过基于时序更新的 ARTrack 0.8%、1.9% 和 1.7%。这表明，采用 DropMAE 预训练的跟踪器能够很好地泛化到通用视觉目标跟踪中的未见物体。

TNL2K。 TNL2K 是一个大规模评估数据集，包含 700 个测试视频，涵盖显著

表 6 在视频目标分割（DAVIS-2017）和人体姿态传播（JHMDB）任务上与先前自监督学习方法的比较。“适配器”表示我们在冻结的骨干网络上添加轻量级适配器，且仅适配器可学习。对于所有 ViT 模型，我们使用大小为 16×16、步幅为 8 的图块。排名最好的两项结果分别以 **红色** 和 **蓝色**标示。

方法	骨干网络	数据集	DAVIS-2017 [13] $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	JHMDB [172] 正确关键点@0.1	正确关键点@0.2
监督式 [110]	ResNet-50	ImageNet	66.0	63.7	68.4	59.2	78.3
TimeCycle [173]	ResNet-50	VLOG	40.7	41.9	28.9	57.7	78.5
UVC [174]	ResNet-18	K400	57.8	56.3	59.2	58.6	79.6
SimSiam [175]	ResNet-50	ImageNet	66.3	64.5	68.2	58.4	77.5
MoCo [45]	ResNet-50	ImageNet	65.4	63.2	67.6	**60.4**	79.3
VINCE [176]	ResNet-50	K400	65.6	63.4	67.8	58.2	76.3
RegionTracker [177]	ResNet-50	TrackingNet	63.4	61.5	65.4	57.5	74.6
CRW [118]	ResNet-18	K400	67.6	64.8	70.2	59.3	80.3
DUL [53]	ResNet-18	YT-VOS	**69.3**	**67.1**	71.6	56.4	79.1
DUL [53]	ResNet-18	K400	68.7	**66.7**	70.7	58.2	80.5
VFS [119]	ResNet-50	K400	68.9	66.5	71.3	**60.9**	**80.7**
MAE [6]	ViT-B/16	ImageNet	59.1	57.1	61.2	-	-
OMNIMAE [178]	ViT-B/16	SSv2+ImageNet	36.2	34.7	37.6	-	-
MME [179]	ViT-B/16	ImageNet	59.2	57.1	61.2	-	-
VideoMAE [54]	ViT-B/16	K400	43.4	41.9	44.9	-	-
DropMAE	ViT-B/16	K400	60.3	58.3	62.3	-	-
VideoMAE [54] + 适配器	ViT-B/16	YT-VOS	57.7	54.9	60.6	-	-
DropMAE +适配器	ViT-B/16	YT-VOS	**69.4**	66.3	**72.5**	57.3	80.2
DropMAE +适配器	ViT-B/16	K400	68.7	65.5	**71.9**	57.8	**80.8**

表 7 与以往自监督方法的训练效率比较。使用冻结的 DropMAE模型，仅微调一个可学习参数较少（5.4M）的适配器，即可达到与 DUL基线相当的性能，同时在 YT-VOS 和 K400 数据集上，训练速度分别提升 4.6× 至 16.6× 倍。最高效率以**粗体**显示。训练时间是在单块 RTX 3090 GPU 上测得的。

方法	学习模型	训练时间	数据集
CRW [118]	ResNet-18 (11.2M)	168 小时	K400
DUL [53]	ResNet-18（11.5M）	182.9 小时	K400
DUL [53]	ResNet-18（11.5M）	16 小时	YT-VOS
DropMAE + 适配器	适配器（5.4M）	**11** 小时	K400
DropMAE + 适配器	适配器（5.4M）	**3.5** 小时	YT-VOS

外观变化和人工添加的对抗样本。如表 2所示，我们的 DropMAE 在该数据集上的表现显著优于其他跟踪器。

ITB、TrackingNet 和 OTB100。 在表3中，我们在 ITB 上评估了 DropTrack [153], OTB100 [152] 和 TrackingNet [150],，在每个数据集上都取得了最先进的性能。DropTrack 在 ITB 和 TrackingNet 上略优于 OSTrack。我们认为，主要原因是这两个数据集的训练和测试对象类别完全重叠，因而降低了预训练的影响。即使使用监督式 ImageNet 权重，也能在这两个数据集上训练出具有竞争力的跟踪器，这一点已在 [3]中展示。

在全部 7 个 VOT 数据集上，我们的 DropTrack 均优于基线 OSTrack，这表明，在视频上进行 DropMAE 预训练所学到的时序匹配表征优于在 ImageNet 上训练的 MAE 模型，从而使跟踪器更加稳健，并能很好地泛化到未见过和已见过的对象。

可视化。 在图 9中，我们展示了 DropTrack 与其他 3 个对比跟踪器获得的定性跟踪结果。所选序列包含各种挑战，包括显著的外观变化、背景聚类、光照变化以及相似对象。得益于稳健的 DropMAE 预训练模型，我们的 DropTrack 能够很好地应对这些挑战。

5.3 视频目标分割

在表 4中，我们将 DropSeg 与现有的 VOS 方-法在 theDAVIS-16/17 [155, 13]上进行了比较。

DAVIS-16。 DAVIS-16 由 20 个经过人工标注的测试序列组成。如表 4所示，我们的单次学习 DropSeg 方法，

表 8 与先前自监督方法在 OTB 上的比较 [152]

100

。最佳结果以**粗体**显示。

方法	骨干网络	数据集	AUC
监督式 [110]	ResNet-50	ImageNet	45.5
SimSiam [175]	ResNet-50	ImageNet	43.2
MoCo [45]	ResNet-50	ImageNet	46.5
VINCE [176]	ResNet-50	K400	47.6
RegionTracker [177]	ResNet-50	TrackingNet	43.4
SeCo [180]	ResNet-50	K400	51.8
VFS [119]	ResNet-50	K400	43.4
VFS [119]	ResNet-50	K400+GOT-10k	52.5
VideoMAE [54] +适配器	ViT-B/16	K400	43.5
DropMAE +适配器	ViT-B/16	K400	**53.2**

表 9 Sintel（训练集）上的光流估计结果。我们在 FlyingChairs [79] 和 FlyingThings3D [169],上训练 DropRAFT，并在 Sintel（训练集）上测试，以评估其泛化性能。

方法	来源	Clean ↓	最终版 ↓
高清3 [181]	CVPR19	3.84	8.77
FlowNet2 [79]	CVPR17	2.02	3.54
PWC-Net [82]	TPAMI19	3.45	4.60
GMA [182]	ICCV21	1.30	2.74
GMFlow [182]	ICCV21	1.08	2.48
SKFlow [89]	NeurIPS22	1.22	2.46
DIP [183]	CVPR22	1.30	2.82
CRAFT [88]	CVPR22	1.27	2.79
RAFT-it [184]	ECCV22	1.74	2.41
GMFlowNet [90]	CVPR22	1.14	2.71
FlowFormer++ [92]	CVPR 2023	0.94	2.33
TransFlow [185]	CVPR 2023	0.93	2.33
EMD-L [186]	ICCV 2023	**0.88**	2.55
RPKNet [187]	AAAI 2024	1.12	2.45
SEA-RAFT（M） [91]	ECCV 2024	1.21	4.04
SEA-RAFT（L） [91]	ECCV 2024	1.19	4.11
RAFT [83]	ECCV20	1.43	2.71
ViT-RAFT	随机	13.56	-
ViT-RAFT	VideoMAE [54]	13.48	-
ViT-RAFT	MAE [6]	1.51	2.72
DropRAFT	Ours	1.06	**2.25**

无需使用任何在线学习和复杂的记忆机制，便取得了最高的 $\mathcal{J}\&\mathcal{F}$ 分数 92.1%，显著优于其他对比的一次性方法，甚至超过了采用复杂流程的方法（即 OL、M 和 S）。这表明，预训练期间学习到的像素级对应关系有助于捕捉 VOS 中不同帧之间的长程依赖关系。

表10 TAP-Vid DAVIS上的长期点跟踪性能480 [95]。我们的测试时自监督跟踪器相较于有监督方法表现良好， [99, 96] 这些有监督方法使用大规模标注数据进行训练；与此同时，我们仅使用更少的可学习参数，便超越了最先进水平的测试时DINO-Tracker。“Param.”表示测试时训练期间可学习的骨干网络参数。模式“S”和“TT”分别表示监督式训练和测试时训练。

方法	模式	参数	骨干网络	δ^x_{avg}	OA	AJ
RAFT [83]		-	CNN	66.7	-	-
DINOv2 [108]		-	ViT-L-14/7	66.7	-	-
TAP-Net [95]	S	-	CNN	66.4	79.0	46.0
PIPs++ [98]	S	-	CNN	73.6	-	-
TAPIR [96]	S	-	CNN	77.3	89.5	65.7
Co-Tracker [99]	S	-	Transf.	79.4	89.5	65.6
OmniMotion [105]	TT	-	MLP	74.1	84.5	58.4
DINO-Tracker [14]	TT	7.59M	CNN	**80.4**	88.1	64.6
DINOv2$^{LoRA-2L}$ [14]	TT	0.07M	ViT-L-14/7	73.2	84.8	58.0
DropDINO$^{LoRA-2L}$	TT	0.05M	ViT-B-16/8	79.0	89.1	64.9
DropDINO$^{LoRA-4L}$	TT	0.10M	ViT-B-16/8	79.5	**89.6**	**65.4**
DropDNO$^{LoRA-6L}$	TT	0.15M	ViT-B-16/8	**79.7**	**89.8**	**65.7**

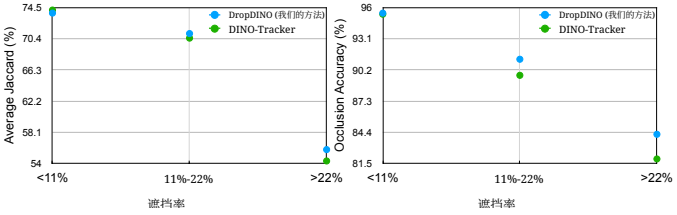

图 11. 不同遮挡率下 TAP-Vid DAVIS-480 [95] 上的跟踪性能。我们根据各视频的平均遮挡率，将 TAP-Vid DAVIS-480 中的测试视频分为三组；平均遮挡率根据真实标注的可见性进行估算。报告平均交并比（AJ）和遮挡准确率（OA）。随着遮挡率增加（即， $>22\%$），我们的 DropDINO 性能下降更少，展现了其在严重遮挡情况下进行长期点跟踪的潜力。

DAVIS-17。 DAVIS-17 是 DAVIS-16 的扩展，包含更具挑战性的视频，并支持多目标分割。在表 4 中，我们的 DropSeg 取得了具有竞争力的结果：83.0%$\mathcal{J}\&\mathcal{F}$、80.2% $\mathcal{J}$ 和 85.7% $\mathcal{F}$，这表明它在处理更具挑战性的视频方面表现出色。

可视化。我们的 DropSeg 定性可视化结果如图 10 所示。即使不使用在线微调或复杂的记忆机制，DropSeg 也能仅凭第一帧的掩码标注，在后续帧中提供准确的分割结果，这主要得益于 DropMAE 学到的良好时序匹配能力。

5.4 3D点云跟踪

在表 5 中，我们评估了 DropMAE 预训练模型在三维点云跟踪中的表现。我们的 DropMAE 变体优于其他基于 Transformer 的 3D 跟踪器（例如 MBPTrack [71] 和 STNet[67]），在成功率和精确度指标方面均表现出色。值得注意的是，DropMAE 的表现甚至可与三维预训练方法（即 Point-MAE [171] 和 Recon [170]）相媲美，这表明从二维视频中学习到的时序匹配能力可以很好地迁移到三维跟踪中。

5.5 自监督对应学习

在表6中，我们将 DropMAE 作为冻结特征提取器，并评估其在 DAVIS-2017 [13] 和 JHMDB [172] 上的无监督视频目标分割和姿态传播性能。在这两项任务中，DropMAE 的表现均优于其他生成模型（例如 MAE、OMNIMAE 和 MME），并显著优于基于视频的生成方法 VideoMAE。这主要是



图 12. DINO-Tracker [14] 和我们的 DropDNO$^{LoRA-6L}$获得的长期点跟踪结果。我们的方法能够更好地处理目标遮挡，并实现更稳健的稠密点跟踪。

因为 VideoMAE 使用 3D卷积神经网络沿时间维度进行立方体提取，但无法学习有效的时序对应关系，因此更适用于高层次的视频动作识别任务。

此外，我们在冻结的 DropMAE 特征提取器之上添加了一个轻量级适配器，并通过[53]进行训练。我们的新变体实现了与 DUL基线 [53], 相当的无监督跟踪性能，同时分别在 YT-VOS 和 K400 上以4.6× 和 16.6× 更快的速度完成训练，如表7所示。值得注意的是，我们在传统的目标级无监督跟踪任务上评估了 DropMAE 学到的表征 + 适配器（见表8），而我们的 DropMAE 实现了领先的性能。

5.6 光流估计

按照标准评估流程， [83], 我们在 FlyingChairs [79] 和 FlyingThings [169]合成数据集上训练 DropRAFT，并在 Sintel 的训练集上测试零样本泛化性能 [93]。如表9所示，我们的 DropRAFT 在 Clean 和 Final子集上均以较大优势显著超越基线 RAFT，这证明了 DropMAE骨干网络在光流估计中的有效性。与使用相同 ViT-B/16 骨干网络的 MAE-RAFT 相比，DropRAFT 获得了更低的平均端点误差，这表明 Drop-MAE 比 MAE 更适合作为时间学习器。FlowFormer++在光流估计中应用了掩码自编码。尽管它是专为光流设计的，我们通用的 DropMAE 仍然在 Final子集上超越了它，展现了 DropMAE 表征在该任务上的泛化能力。对于采用 Random 和 VideoMAE 初始化的 ViT-RAFT，我们观察到它们遭遇了严重的训练崩溃。鉴于 MAE 和我们的 DropRAFT 所取得的成功，预训练权重对于基于ViT的光流估计至关重要。

5.7 长期点跟踪

如§3.4.5所示，我们在 DINO-Tracker 框架中使用 DropMAE 作为特征提取器。为了进行参数高效训练，我们采用 LoRA 训练，专门优化骨干网络的最后 N 层（即表 NL 中的10）。从表10可以看出，我们的 DropDINO$^{LoRA-2L}$仅优化 DropMAE 的最后两层，且可学习参数更少（0.07M），但在遮挡准确率（OA）和平均交并比（AJ）两项指标上均优于基于测试时优化的最先进 DINO-Tracker [14] 。在点准确率（δ^x_{avg}）指标上，即使使用图块大小较大的粗粒度 ViT 补丁嵌入层，我们的 DropDINO 也取得了与 DINO-Tracker 相当的性能

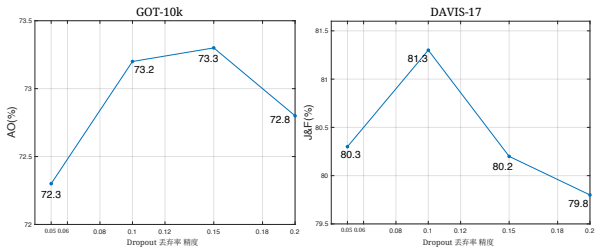

图 13. DropMAE 中丢弃率 P 对 GOT-10k（VOT）和 DAVIS-17（视频目标分割）数据集的消融研究。

表 11. 使用在不同视频数据集上预训练的 DropMAE 所获得的 GOT-10k 和 DAVIS-17 上的下游 VOT 和视频目标分割性能。视频目标分割使用 800 个预训练周期。

数据集	No. 视频	No. 动作	VOT 平均重叠率 $SR_{0.5}$	VOT $SR_{0.75}$	VOS $\mathcal{J}\&\mathcal{F}$
K400 [144]	240,000	400	73.2	83.9 67.5	82.7
K600 [145]	390,000	600	74.5	85.5 69.5	82.8
K700 [146]	526,768	700	**75.6**	**86.2** **71.4**	**83.0**
MiT [147]	802,244	339	75.1	85.5 70.6	82.8
WebVid [148]	240,000	-	72.8	83.4 67.3	81.5
WebVid [148]	960,000	-	73.4	85.0 69.5	82.9

图块大小为 16×16，而 DINO-Tracker 使用粒度更细的 CNN 卷积核，大小为 7×7。在 DropDINO 中增加 LoRA 训练的层数，可以持续提升各项指标，同时仅小幅增加参数数量。此外，[14] 使用 DINOv2 作为 LoRA 训练的骨干网络（记为 DINOv2$^{LoRA-2L\dagger}$），但性能有所下降，这主要是由于 DINOv2 缺乏时间先验。与使用大规模数据集进行监督式训练的方法（例如 Co-Tracker [99]和 TAPIR [96]）相比，我们的 DropDINO$^{LoRA-6L}$ 使用更少的可学习参数，却取得了相当的性能（见图11），展现出作为长时稠密点跟踪通用骨干网络的潜力。

可视化。 图 12 展示了我们的 DropDINO 和 DINO-Tracker 的定性点跟踪结果。12我们的方法能更有效地处理目标遮挡（见图 11），并准确跟踪密集点，这主要得益于 DropDINO 中高效的时间学习器（即 DropMAE）。

6 消融研究

本节通过消融研究对我们的方法进行更详细的分析。我们使用经过 400 个 epoch 预训练的 DropMAE 进行消融研究。

丢弃率的影响 P。我们在图 P 中研究了丢弃率的影响。13。较小的丢弃率 $P = 0.1$ 在 VOT 和视频目标分割任务中均表现良好。同时，丢弃过多空间线索（例如 P=0.2）会降低下游任务的表现，这主要是因为空间线索也有助于实现精确定位和分割。$P = 0.1$ 是最优设置，因此我们在后续实验中采用该设置。

预训练视频来源。由于我们是首个探索用于时序匹配任务的掩码自编码器预训练的研究，因此尚不清楚哪种视频数据集最适合用于预训练。此处，我们使用五种常见的视频数据集进行预训练，包括 K400 [144], K600 [145], K700 [146], MiT [147] 和 WebVid- 2M [148]。对于 WebVid-2M，为了公平比较并加快验证速度，我们随机抽取了 24 万和 96 万个视频。下游跟踪结果见表11。即使使用 WebVid 中的 96 万个视频进行预训练，性能仍不理想。

表 12 不同设置的变体在 GOT10-k 上报告的 AO/$SR_{0.75}$ 跟踪性能。

设置	GOT-10k
DropTrack-K400-400E	73.2/67.5
编码器中使用 ASAD	73.1/68.1
使用领域特定数据	73.4/68.8
不使用帧身份嵌入	72.9/67.4

表 13 最大采样帧间隔对下游跟踪任务的影响。

最大帧采样间隔	GOT-10k AUC	$SR_{0.5}$	$SR_{0.75}$
1	72.2	82.7	65.7
10	72.8	83.4	67.2
50	**73.2**	**83.9**	**67.5**

这表明 WebViD 并非适合跟踪预训练的选择，主要是因为它是一个视频描述数据集，侧重于场景多样性，却缺乏丰富的物体运动信息。使用 K400/600/700 或 MiT 时，跟踪任务能够受益于基于丰富动作类别（即 K700 的 700 个动作类别）的预训练，模型可借此学习到更强的时序匹配能力。

在编码器中应用 ASAD。我们测试了在掩码自动编码器的编码器和解码器的所有层中应用 ASAD。如表12所示，该变体优于原始基线（SR 提高了 0.6%）$_{0.75}$。考虑到其额外开销和有限的提升，我们仅在解码器中应用 ASAD。

领域特定数据。我们还将跟踪训练数据（不使用边界框标注），包括 TrackingNet、LaSOT 和 GOT-10k，加入 K400 进行预训练。使用更大预训练集时，下游跟踪性能为 73.4/68.8，优于基线。这表明领域特定数据有助于弥合领域差距；未来可以考虑通过纳入更多跟踪视频来扩展 Kinetic 数据集。

帧身份嵌入。在预训练期间，帧身份嵌入用于识别两帧中二维位置相同的掩码图像块。从表12可以看出，下游微调若不使用帧身份嵌入，效果会比使用时差，因为不使用它与预训练阶段的做法不一致。

最大帧采样间隔的影响。在预训练期间，我们从一个训练视频中随机采样两帧，并预先设定最大帧采样间隔 g。在这里，我们研究其对下游 VOT 任务的影响。如表 13所示，较大的帧采样间隔（即 50）更有利于 VOT 任务。这是因为使用相对较大的帧采样间隔可以学到更强的时序匹配能力。鉴于从 $g = 10$ 到 $g = 50$的性能提升有限，我们在所有预训练实验中直接使用 $g = 50$。

从 K400 学习静态帧表征。为了证明预训练中的时序对应学习是下游跟踪任务取得成功的关键，我们将 K400[144] 视为静态图像数据集，并在其上执行原始 MAE 预训练，记为 MAE-K400-static。具体而言，在每次训练迭代中，随机从一个视频中采样一帧图像用于掩码自编码预训练。为了与 DropMAE 公平比较，我们将该基线中的视频数量加倍，使每轮训练采样的帧总数与 DropMAE 相同。

MAE-K400-static 与 DropMAE 的比较结果如表 14所示。缺少时序对应学习时，MAE-K400-static 的表现显著差于 DropMAE，这进一步证

表 14 DropMAE、MAE-K400-static 和 RandDrop-MAE 在 GOT-10k 上的比较 [112]。

方法	GOT-10k		
	AUC	$SR_{0.5}$	$SR_{0.75}$
DropMAE	**73.2**	**83.9**	**67.5**
MAE-K400-static	70.4	80.7	65.6
RandDrop-MAE	71.7	82.4	66.2

表15 使用预训练 MAE 模型初始化 DropMAE 预训练的消融研究。

预训练MAE	GOT-10k			Davis-17
	AUC	$SR_{0.5}$	$SR_{0.75}$	$\mathcal{J}\&\mathcal{F}$
w/o	73.2	83.9	67.5	81.3
w/	**75.2**	**85.4**	**71.5**	**82.6**

明了 DropMAE 预训练中时序对应学习的有效性。

随机丢弃。原始ViT [32] 实现了丢弃 [188]机制，并将其应用于每个多头自注意力层。为考察这种随机丢弃在我们的掩码自编码预训练设置中是否有效，我们构建了一个名为 RandDrop-MAE 的基线，在预训练期间于解码器的每个自注意力层中采用随机丢弃。与我们的自适应丢弃策略（即 ASAD）不同，RandDrop-MAE 会随机丢弃帧间或帧内注意力。为进行公平比较，我们为 RandDrop-MAE 使用相同的丢弃率（即 0.1）。如表14所示，与我们的 DropMAE 相比，RandDrop-MAE 的性能有所下降。我们认为，这是因为随机丢弃可能会过度丢弃一些对重建至关重要的注意力元素，从而降低学习效果。

预训练MAE。下游 VOT 和 VOS 任务包含大量类别各异的目标用于评估。考虑到 K400 由人类动作视频组成，预训练与微调阶段之间仍存在领域差距。为缩小这一差距，我们使用在 ImageNet 上训练的原始 MAE 作为 DropMAE 的预训练权重，随后进一步在 K400 上对 DropMAE 进行预训练，以学习时序对应关系。从表15可见，我们的 DropMAE 在 VOT 和 VOS 任务中均受益于预训练MAE，这主要是因为 MAE 学到的多样化目标类别有助于通用目标跟踪和分割。这也表明，使用预训练MAE和规模更大的视频数据源（例如 K700 [146]）有望提升下游任务性能。

帧重建可视化。图 14 展示了我们提出的 DropMAE 所获得的视频帧重建结果。14尽管重建过程中利用的空间线索较少，我们的 DropMAE 仍通过探索时间线索或帧间图像块，取得了良好的重建结果，证明了所提方法的有效性。

7 结论

本文研究了面向多种基于时序匹配的下游任务的掩码自编码预训练，包括目标级跟踪（即 VOT 和视频目标分割）、像素级跟踪（即光流估计和长期点跟踪）、三维点云跟踪以及自监督视觉对应关系学习。具体而言，我们提出一种自适应空间注意力丢弃方法，以促进从二维视频中学习时间对应关系。值得注意的是，我们发现 DropMAE 的下游跟踪性能优于基于图像的 MAE，同时50% 缩短了预训练时间。我们在



图 14. DropMAE 在 K400 验证集上的视频帧重建结果。依次展示原始输入帧对、掩码帧对（即掩码比例为 75%）和重建结果。

个下游跟踪任务和 13 项基准测试上的实验表明，DropMAE 在各种跟踪应用中的有效性。我们希望 DropMAE 能成为适用于各种跟踪任务的通用预训练骨干网络，并推动跟踪领域开展更多预训练研究。未来的工作将探索把 DropMAE 扩展到各种模型架构（例如，状态空间模型）、跟踪范式（例如，自回归跟踪），以及更高效的预训练策略。

8 致谢

本研究获得中国香港特别行政区研究资助局（项目编号：CityU 11215820）和香港城市大学（项目编号：7030010）的资助。

参考文献

[1] T. Brown、B. Mann、N. Ryder、M. Subbiah、J. D. Kaplan、P. Dhariwal 和 A. Neelakantan，《语言模型是少样本学习者》，发表于神经信息处理系统大会，2020 年。[2] A. Devlin 、 M.-W. Chang 、K. Lee、和 K. Toutanova，《语言模型是少样本学习者》，发表于 *NAACL*，2019年。[3] B. Ye、H. Chang、B. Ma 和 S. Shan，《用于跟踪的联合特征学习与关系建模：单流框架》，发表于 欧洲计算机视觉会议，2022年，第341–357页。[4] T. Truong 和 A. Pham，《畅所欲言：一种音视频 Transformer 方法》，发表于 *CVPR*，2021年，第1105–1114页。[5] N. Li、S. Liu 和 Y. Liu，《基于 Transformer 网络的神经语音合成》，发表于 *AAAI*，2019年，第6706–6713页。[6] K. He、X. Chen 和 S. Xie，《掩码自编码器是可扩展的视觉学习器》，发表于 *CVPR*，2022年，第16 000–16 009页。[7] Z. Xie、Z. Zhang、Y. Cao、Y. Lin、J. Bao 和 Z. Yao，《SimMIM：一种用于掩码图像建模的简单框架》，发表于 *CVPR*，2022年，第9653–9663页。[8] B. Chen、P. Li、L. Bai、L. Qiao、Q. Shen 和 B. Li，《骨干网络即所需：一种用于视觉目标跟踪的简化架构》，发表于欧洲计算机视觉会议，2022年。[9] O. Russakovsky、J. Deng 等，《ImageNet 大规模视觉识别挑战》，发表于国际计算机视觉杂志，第115卷，第3期，第211–252页，2015年。[10] Q. Wu、J. Wan 和 A. B. Chan，《用于视觉目标跟踪的渐进式无监督学习》，发表于 *CVPR*，2021年。[11] L. Bertinetto、J. Valmadre、J. Henriques、A. Vedaldi 和 P. Vedaldi，《用于目标跟踪的全卷积孪生网络》，发表于 *ECCVW*，2016，第850–865页。[12] B. Li、W. Wu、Z. Zhu 和 J. Yan，《使用孪生区域提议网络实现高性能视觉跟踪》，发表于 *CVPR* 论文集，2018年，第8971–8980页。

[13] J. Pont-Tuset、F. Perazzi、S. Caelles、P. Arbeláez、A. Sorkine-Hornung 和 L. Van Gool，“2017年视频目标分割 DAVIS 挑战赛”，收录于*arXiv:1704.00675*，2017年。[14] N. Tumanyan、A. Singer、S. Bagon 和 T. Dekel，“DINO-Tracker：在单个视频中驯服 DINO 以进行自监督点跟踪”，收录于欧洲计算机视觉会议，2025年，第 367–385 页。[15] Q. Wu、T. Yang、Z. Liu、B. Wu、Y. Shan 和 A. B. Chan，“DropMAE：用于跟踪任务的空间注意力丢弃掩码自编码器”，收录于 CVPR，2023年，第 14 561–14 571 页。[16] Q. Wu 和 A. Chan，“面向视觉目标跟踪的元图适应”，收录于国际多媒体与博览会会议，2021。[17] Y. Liu、Y. Liang、Q. Wu、L. Zhang 和 H. Wang，“一种基于多重深度相关滤波器的目标跟踪新框架”，收录于 IEEE声学、语音与信号处理国际会议，2022。[18] H. Galoogahi、A. Fagg 和 S. Lucey，“学习用于视觉跟踪的背景感知相关滤波器”，收录于 IEEE国际计算机视觉会议，2017。[19] J. F. Henriques、R. Caseiro、P. Martins 和 J. Batista，“利用核化相关滤波器实现高速跟踪”，发表于 IEEE模式分析与机器智能汇刊，第 37 卷，第 3 期，第 583–596 页，2015年。[20] Y. Liang、Q. Wu 和 H. Wang，“ 利用引导式实例感知提议进行深度相关滤波跟踪 ”，收录于《IEEE智能交通系统汇刊》，2021。[21] X. Wang、C. Li、B. Luo 和 J. Tang，“Sint++：通过对抗式正样本生成实现鲁棒视觉跟踪”，收录于 CVPR，2018，第 4864–4873 页。[22] B. Li、W. Wu、Q. Wang、F. Zhang、J. Xing 和 J. Yan，“SiamRPN++：基于超深网络的孪生视觉跟踪演进”，收录于 CVPR，2019。[23] T. Yang 和 A. B. Chan，“学习用于目标跟踪的动态记忆网络”，收录于欧洲计算机视觉会议，2018，第 152–167 页。[24] L. Zhang、J. Weijer、M. Danelljan 和 F. Khan，“学习孪生跟踪器的模型更新”，收录于 IEEE国际计算机视觉会议，2019。[25] H. Fan 和 H. Ling，“用于实时视觉跟踪的孪生级联区域提议网络”，收录于 IEEE国际计算机视觉会议，2019。[26] Z. Zhang 和 H. Peng，“用于实时视觉跟踪的更深更宽的孪生网络”，收录于 CVPR，2017。[27] S. W. Oh、J. Y. Lee、N. Xu 和 S. J. Kim，“利用时空记忆网络进行视频目标分割”，收录于 IEEE国际计算机视觉会议，2019，第 9226–9235 页。[28] Z. Yang、Y. Wei 和 Y. Yang，“利用变换器关联视频目标分割中的对象”，收录于神经信息处理系统大会，2021。[29] H. K. Cheng、Y. W. Tai 和 C. K. Tang，“重新思考时空网络，以改进记忆覆盖率并实现高效视频目标分割”，收录于神经信息处理系统大会，2021。[30] H. K. Cheng 和 A. G. Schwing，“XMem：基于阿特金森-希夫林记忆模型的长时视频目标分割”，收录于欧洲计算机视觉会议，2022。[31] Z. Lin、T. Yang、M. Li、Z. Wang、C. Yuan、W. Jiang 和 W. Liu，“SWEM：利用序贯加权期望最大化实现实时视频目标分割”，收录于 CVPR，2022，第 1362–1372 页。[32] A. Dosovitskiy、L. Beyer 和 A. Kolesnikov，“一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer”，收录于国际学习表征会议，2021。[33] W. Cai、Q. Liu 和 Y. Wang，“HIPTrack：利用历史提示进行视觉跟踪”，收录于 CVPR，2024。[34] F. Xie、Z. Wang 和 C. Ma，“DiffusionTrack：用于视觉目标跟踪的点集扩散模型”，收录于 CVPR，2024年6月。[35] J. Shi、Y. Yu、B. Hui、J. Shi 和 H. Luo，“ 用于视觉跟踪的历史状态建模 ”，《神经计算与应用》，第 1–18 页，2025。[36] C. Alawode、S. Javed、A. Mahmood 和 J. Matas，“预测表现最佳的视觉跟踪器”，收录于 arXiv:2407.15707，2024。[37] R. Zhang、P. Isola 和 A. A. Efros，“图像着色”，收录于欧洲计算机视觉会议，2016。[38] M. Noroozi 和 P. Favaro，“通过解拼图进行视觉表征的无监督学习”，收录于欧洲计算机视觉会议，2016。[39] N. Srivastava 和 R. Salakhudinov，“利用长短期记忆网络进行视频表征的无监督学习”，收录于国际机器学习会议，2015。[40] C. Vondrick、H. Pirsiavash 和 A. Torralba，“生成具有场景动态的视频”，载于神经信息处理系统大会，2016年。[41] S. Gidaris、P. Singh 和 N. Komodakis，“通过预测图像旋转进行无监督表征学习”，载于 arXiv:1803.07728，2018年。[42] Y. Cao、Z. Xie、B. Liu、Y. Lin、Z. Zhang 和 H. Hu，“用于无监督视觉特征学习的非参数实例分类”，载于神经信息处理系统大会，2020年。[43] T. Chen 和 M. Norouzi，“用于视觉表征对比学习的简单框架”，载于国际机器学习会议，2020年。[44] A. Dosovitskiy、J. Tobias、M. Riedmiller 和 T. Brox，“利用卷积神经网络进行判别式无监督特征学习”，载于神经信息处理系统大会，2014年，第766–774页。[45] K. He、H. F. adn Y. Wu，“用于无监督视觉表征学习的动量对比”，载于CVPR，2020年。[46] Z. Wu、Y. Xiong 和 S. Yu，“通过非参数实例判别进行无监督特征学习”，载于CVPR，2018年。[47] Z. Xie、Y. Lin、Z. Zhang、Y. Cao、S. Lin 和 H. Hu，“传播自我：探索无监督视觉表征学习中的像素级一致性”，载于CVPR，2021年。[48] C. Feichtenhofer、H. Fan、B. Xiong、R. Girshick 和 K. He，“无监督时空表征学习的大规模研究”，载于CVPR，2021年，第3299–3309页。[49] Y. Bai、H. Fan、Y. Lu、Y. Zhou、Q. Yu、V. Chandra 和 A. Yuille，“时间信息能否助力对比式自监督学习？ ” arXiv:2011.13046，2020年。[50] I. Dave、R. Gupta、M. N. Rizve 和 M. Shah，“Tclr：用于视频表征的时间对比学习”，《计算机视觉与图像理解》，第219卷，2022年。[51] T. Han、W. Xie 和 A. Zisserman，“用于视频表征学习的自监督协同训练”，神经信息处理系统大会，第33卷，第5679–5690页，2020年。[52] R. Qian、T. Meng、B. Gong、M.-H. Yang、H. Wang、S. Belongie 和 Y. Cui，“时空对比视频表征学习”，载于CVPR，2021年，第6964–6974页。[53] N. Araslanov 和 S. Roth，“用于视频分割的密集无监督学习”，神经信息处理系统大会，第34卷，25 319，2021年，25 308页。[54] Z. Tong、Y. Song、J. Wang 和 L. Wang，“VideoMAE：掩码自编码器是适用于自监督视频预训练的数据高效学习器”，载于神经信息处理系统大会，2022年。[55] C. Feichtenhofer、H. Fan、Y. Li 和 K. He，“作为时空学习器的掩码自编码器”，载于神经信息处理系统大会，2022年。[56] A. Gupta、J. Wu、J. Deng 和 F.-F. Li，“孪生掩码自编码器”，神经信息处理系统大会，第36卷，第40 676–40 693页，2023年。[57] H. Qi、C. Feng、Z. Cao、F. Zhao 和 Y. Xiao，“P2B：用于点云中三维目标跟踪的点到框网络”，载于CVPR，2020年，第6329–6338页。[58] A. Geiger、P. Lenz、C. Stiller 和 R. Urtasun，“视觉与机器人学的交汇：KITTI 数据集”，《国际机器人研究杂志》，第32卷，第11期，第1231–1237页，2013年。[59] H. Caesar、V. Bankiti、A. H. Lang、S. Vora、V. E. Liong、Q. Xu、A. Krishnan、Y. Pan、G. Baldan 和 O. Beijbom，“nuScenes：自动驾驶多模态数据集”，载于CVPR，2020年。[60] P. Sun、H. Kretzschmar、J. Guo、Y. Zhou、Y. Chai 和 B. Caine，“自动驾驶感知的可扩展性：Waymo开放数据集”，载于CVPR，2020年，第2446–2454页。[61] S. Giancola、J. Zarzar 和 B. Ghanem，“利用形状补全进行三维孪生跟踪”，载于CVPR，2019年。[62] C. Zheng、X. Yan、J. Gao、W. Zhao、W. Zhang、Z. Li 和 S. Cui，“面向点云单目标跟踪的框感知特征增强”，载于IEEE国际计算机视觉会议，2021年，第13 199–13 208页。[63] L. Hui、L. Wang、M. Cheng、J. Xie 和 J. Yang，“用于稀疏点云的三维孪生体素到鸟瞰图跟踪器”，神经信息处理系统大会，第34卷，第28 714–28 727页，2021年。[64] Y. Cui、Z. Fang、J. Shan、Z. Gu 和 S. Zhou，“基于Transformer的三维目标跟踪”，arXiv:2110.14921，2021年。[65] C. Zhou、Z. Luo、Y. Luo、T. Liu、L. Pan、Z. Cai、H. Zhao 和 S. Lu，“PTTR：基于Transformer的关系式三维点云目标跟踪”，载于CVPR，2022年，第8531–8540页。[66] Z. Guo, Y. Mao, W. Zhou, M. Wang, 和 H. Li，“Cmt：上下文-

匹配 g-g用于三维跟踪的引导式 Transformer g in p点云 ，”发表于欧洲计算机视觉会议，2022年，第95–111页。[67] L. Hui、L. Wang、L. Tang、K. Lan、J. Xie和J. Yang，“用于点云单目标跟踪的三维孪生变换器网络”，发表于欧洲计算机视觉会议，2022年，第293–310页。[68] S. Feng、J. Gao和E. Cheng，“采用密集连接的多相关孪生变换器网络，用于三维单目标跟踪”，IEEE机器人与自动化快报，2023年。[69] Q. Wu、C. Sun和J. Wang，“用于稀疏点云三维单目标跟踪的多层结构增强网络”，机器人与自动化快报，第8卷，第1期，第9–16页，2022年。[70] T.-X. Xu、Y.-C. Guo、Y.-K. Lai和S.-H. Zhang，“Cxtrack：利用上下文信息改进三维点云跟踪”，发表于CVPR，2023年，第1084–1093页。[71] T.-X. Xu、Y.-C. Guo和Y.-K. Lai，“MBPTrack：利用记忆网络和边界框先验改进三维点云跟踪”，发表于IEEE国际计算机视觉会议，2023年，第9911–9920页。[72] K. Lan、H. Jiang和J. Xie，“时间感知孪生跟踪器：将时间上下文融入三维目标跟踪”，发表于亚洲计算机视觉会议，2022年，第399–414页。[73] C. Zheng、X. Yan、H. Zhang、B. Wang、S. Cheng、S. Cui和Z. Li，“超越三维孪生跟踪：一种面向点云三维单目标跟踪的以运动为中心的范式”，发表于CVPR，2022年，第8111–8120页。[74] Y. Xia、Q. Wu、W. Li、A. B. Chan和U. Stilla，“一种轻量级、无检测器的点云三维单目标跟踪器”，IEEE智能交通系统汇刊，2023年。[75] Q. Wu、Y. Xia、J. Wan和A. B. Chan，“通过二维匹配蒸馏和三维预训练提升三维单目标跟踪”，2023年。[76] Q. Chen和V. Koltun，“Full Flow：通过在规则网格上进行全局优化估计光流”，发表于CVPR，2016年。[77] C. Zach、T. Pock和H. Bischof，“一种用于实时 TV-L1 光流的基于对偶的方法”，发表于《模式识别》，2007年，第214–223页。[78] B. K. Horn和B. G. Schunck，“确定光流。《人工智能》”，1981年。[79] A. Dosovitskiy、P. Fischer、E. Ilg、P. Hausser、C. Hazirbas、V. Golkov、P. Van Der Smagt、D. Cremers和T. Brox，“FlowNet：利用卷积网络学习光流”，发表于IEEE国际计算机视觉会议，2015年，第2758–2766页。[80] E. Ilg、N. Mayer、T. Saikia、M. Keuper、A. Dosovitskiy和T. Brox，“FlowNet 2.0：深度网络光流估计的发展”，发表于CVPR，2017年，第2462–2470页。[81] J. Xu、R. Ranftl和V. Koltun，“通过直接代价体处理实现精确光流”，发表于CVPR，2017年，第1289–1297页。[82] D. Sun、X. Yang、M.-Y. Liu和J. Kautz，“PWC-Net：利用金字塔、扭曲和代价体进行光流估计的卷积神经网络”，发表于CVPR，2018年，第8934–8943页。[83] Z. Teed和J. Deng，“RAFT：用于光流的循环全对场变换”，发表于欧洲计算机视觉会议，2020年，第402–419页。[84] Z. Huang、X. Shi、C. Zhang、Q. Wang、K. C. Cheung、H. Qin、J. Dai和H. Li，“FlowFormer：用于光流的变换器架构”，发表于欧洲计算机视觉会议，2022年，第668–685页。[85] A. Jahedi、M. Luz、M. Rivinius、L. Mehl和A. Bruhn，“MS-RAFT+：高分辨率多尺度RAFT”，《国际计算机视觉杂志》，第132卷，第5期，第1835–1856页，2024年。[86] A. Luo、F. Yang、X. Li和S. Liu，“利用核补丁注意力学习光流”，发表于CVPR，2022年，第8906–8915页。[87] A. Luo、F. Yang、X. Li、L. Nie、C. Lin、H. Fan和S. Liu，“GaFlow：将高斯注意力融入光流”，发表于IEEE国际计算机视觉会议，2023年，第9642–9651页。[88] X. Sui、S. Li、X. Geng、Y. Wu、X. Xu、Y. Liu、R. Goh和H. Zhu，“CRAFT：用于鲁棒光流的交叉注意力光流变换器”，发表于CVPR，2022年，第17 602–17 611页。[89] S. Sun、Y. Chen、Y. Zhu、G. Guo和G. Li，“SKFlow：利用超核学习光流”，神经信息处理系统大会，第35卷，2022年。[90] S. Zhao、L. Zhao、Z. Zhang、E. Zhou和D. Metaxas，“利用重叠注意力进行全局匹配以估计光流tion，”发表于 *CVPR* ，2022页码。17 592–17 601。[91] D. Teng、S. Liao、A. Ranjan、A. Kuzmin、K. Deng、L. Luo、L. Lepson、J. Yang、S. Gupta、B. Ben-Baruch、I. Kligvasser、R. Ronen、N. Kiryati、M. Leordeanu-Kligler、O. Shocher、E. Ben-Baruch、A. Noy、以及 L. Zelnik-Manor，“SEA-RAFT：用于光流的简单、高效、准确的 RAFT”，载于欧洲计算机视觉会议，2025年，第 36–54 页。[92] X. Shi、Z. Huang、D. Li、M. Zhang、K. C. Cheung、S. See、H. Qin、J. Dai 和 H. Li，“FlowFormer++：用于预训练光流估计的掩码代价体自动编码 ”，载于 CVPR，2023年，第 1599–1610 页。[93] D. J. Butler、J. Wulff、G. B. Stanley 和 M. J. Black，“用于光流评估的自然场景开源影片”，载于欧洲计算机视觉会议，2012年，第 611–625 页。[94] A. W. Harley、Z. Fang 和 K. Fragkiadaki，“Particle video 再探：利用点轨迹跟踪遮挡”，载于欧洲计算机视觉会议，2022年，第 59–75 页。[95] C. Doersch、A. Gupta、L. Markeeva、A. Recasens、L. Smaira、Y. Aytar、J. Carreira、A. Zisserman 和 Y. Yang，“TAP-Vid：视频中任意点跟踪的基准”，神经信息处理系统大会，第 35 卷，第 13 610–13 626 页，2022年。[96] C. Doersch、Y. Yang、M. Vecerik、D. Gokay、A. Gupta、Y. Aytar、J. Carreira 和 A. Zisserman，“TAPIR：利用逐帧初始化和时间细化跟踪任意点”，载于 IEEE国际计算机视觉会议，2023年，第 10 061–10 072 页。[97] M. Neoral、J. ˇSer`ych 和 J. Matas，“MFT：每个像素的长期跟踪”，载于 IEEE/CVF冬季计算机视觉应用会议论文集，2024年，第 6837–6847 页。[98] Y. Zheng、A. W. Harley、B. Shen、G. Wetzstein 和 L. J. Guibas，“PointOdyssey：用于长期点跟踪的大规模合成数据集”，载于 IEEE国际计算机视觉会议，2023年，第 19 855–19 865 页。[99] N. Karaev、I. Rocco、B. Graham、N. Neverova、A. Vedaldi 和 C. Rupprecht，“Cotracker：协同跟踪效果更佳”，载于欧洲计算机视觉会议，2025年，第 18–35 页。[100] N. Karaev、I. Makarov、J. Wang、N. Neverova、A. Vedaldi 和 C. Rupprecht，“Cotracker3：通过对真实视频进行伪标签生成实现更简单、更优的点跟踪”，arXiv 预印本 arXiv:2410.11831，2024年。[101] H. Li、H. Zhang、S. Liu、Z. Zeng、T. Ren、F. Li 和 L. Zhang，“Taptr：将变换器作为检测器来跟踪任意点”，载于欧洲计算机视觉会议，2025年，第 57–75 页。[102] S. Cho、J. Huang、J. Nam、H. An、S. Kim 和 J.-Y. Lee，“用于点跟踪的局部全对对应关系”，载于欧洲计算机视觉会议，2025年，第 306–325 页。[103] A. Shrivastava 和 A. Owens，“通过对比随机游走进行自监督任意点跟踪”，arXiv:2409.16288，2024年。[104] G. Aydemir、W. Xie 和 F. G¨uney，“视觉基础模型能否实现长期点跟踪？”，arXiv:2408 13575，2024年。。[105] Q. Wang、Y.-Y. Chang、R. Cai、Z. Li、B. Hariharan、A. Holynski 和 N. Snavely，“Tracking everything everywhere all at once”，载于 IEEE国际计算机视觉会议，2023年，第 19 795–19 806 页。[106] R. Li 和 D. Liu，“分解让 Tracking everything everywhere all at once 更优”，载于欧洲计算机视觉会议，2025年，第 220–235 页。[107] Y. Song、J. Lei、Z. Wang、L. Liu 和 K. Daniilidis，“快速且稳健地跟踪一切、无处不在”，载于欧洲计算机视觉会议，2025年。[108] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等，“DINOv2：无需监督即可学习稳健视觉特征”，arXiv:2304.07193，2023年。[109] Y. Li、H. Mao、R. Girshick 和 K. He，“探索用于目标检测的纯视觉变换器骨干网络”，载于欧洲计算机视觉会议，2022年。[110] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 CVPR，2016年，第 770–778 页。[111] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang、W. Chen 等，“LoRA：大语言模型的低秩适配”。国际学习表征会议，2022年。[112] L. Huang、X. Zhao 和 K. Huang，“GOT-10k：野外通用目标跟踪的大规模高多样性基准”，IEEE模式分析与机器智能汇刊，2019年。[113] H. Touvron、M. Cord、M. Douze、F. Massa、A. Sablayrolles 和 H. J´egou，“通过注意力蒸馏训练数据高效的图像变换器”，载于国际机器学习会议，2021年。[114] T. Ridnik

Manor，“面向大众的 ImageNet-21k 预训练”，载于 arXiv:2104.10972，2021。[115] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal 和 I. Sutskever，“通过自然语言监督学习可迁移视觉模型”，载于国际机器学习会议，2021。[116] X. Chen、S. Xie 和 K. He，“自监督视觉变换器训练的实证研究”，载于 IEEE 国际计算机视觉会议，2021。[117] H. Bao、L. Dong 和 F. Wei，“BeiT：图像变换器的 BERT 预训练”，载于 arXiv:2106.08254，2021。[118] A. Jabri、A. Owens 和 A. Efros，“作为对比随机游走的时空对应关系”，载于神经信息处理系统大会，2020。[119] J. Xu 和 X. Wang，“重新思考自监督对应关系学习：视频帧级相似性视角”，载于 IEEE 国际计算机视觉会议，2021，第 10 075–10 085 页。[120] Z. Wang、H. Zhao、Y. Li、S. Wang、P. Torr 和 L. Bertinetto，“ 不同的跟踪任务是否需要不同的外观模型？ ”，载于神经信息处理系统大会，2021。[121] X. Wang、X. Shu、Z. Zhang、Y. Wang、Y. Tian 和 F. Wu，“借助自然语言实现更灵活、准确的目标跟踪：算法与基准”，载于 CVPR，2021。[122] H. Fan、H. Bai、L. Lin、F. Yang、P. Chu、G. Deng、S. Yu、M. Huang、J. Liu 和 Y. Xu，“LaSOT：高质量、大规模单目标跟踪基准”，载于《国际计算机视觉杂志》，2021。[123] H. Fan、L. Lin 和 F. Yang，“LaSOT：大规模单目标跟踪的高质量基准”，载于 CVPR，2019。[124] H. Nam 和 B. Han，“学习用于视觉跟踪的多域卷积神经网络”，载于 CVPR，2016。[125] M. Danelljan、G. Bhat、F. S. Khan 和 M. Felsberg，“ECO：用于跟踪的高效卷积算子”，载于 CVPR，2017。[126] G. Bhat、M. Danelljan、L. V. Gool 和 R. Timofte，“学习用于跟踪的判别式模型预测”，载于 IEEE 国际计算机视觉会议，2019。[127] P. Voigtlaender、J. Luiten、P. Torr 和 B. Leibe，“Siam R-CNN：通过重新检测实现视觉跟踪”，载于 CVPR，2020。[128] K. Dai、Y. Zhang、D. Wang、H. Lu 和 X. Yang，“借助元更新器实现高性能长期跟踪”，载于 CVPR，2020。[129] Z. Zhang、H. Peng、J. Fu、B. Li 和 W. Hu，“ Ocean：目标感知的无锚框跟踪”，载于欧洲计算机视觉会议，2020。[130] N. Wang、W. Zhou、J. Wang 和 H. Li，“Transformer 遇上跟踪器：利用时间上下文实现鲁棒视觉跟踪”，载于 IEEE 国际计算机视觉会议，2021。[131] Z. Zhang、Y. Liu、X. Wang、B. Li 和 W. Hu，“学习匹配：视觉跟踪的自动匹配网络设计”，载于 IEEE 国际计算机视觉会议，2021，第 19 339–13 348 页。[132] B. Yan、H. Peng、J. Fu、D. Wang 和 H. Lu，“学习用于视觉跟踪的时空 Transformer”，载于 IEEE 国际计算机视觉会议，2021。[133] C. Mayer、M. Danelljan、D. Paudel 和 L. Van Gool，“学习目标候选关联，以持续跟踪不应跟踪的内容”，载于 IEEE 国际计算机视觉会议，2021。[134] Y. Cui、C. Jiang、L. Wang 和 G. Wu，“MixFormer：通过迭代混合注意力实现端到端跟踪”，载于 CVPR，2022。[135] D. Zhang、Y. Fu 和 Z. Zheng，“UAST：不确定性感知孪生跟踪”，载于国际机器学习会议，2022。[136] S. Gao、C. Zhou、C. Ma、X. Wang 和 J. Yuan，“Transformer 视觉跟踪中的注意力之上再加注意力”，载于欧洲计算机视觉会议，2022。[137] Z. Pi、W. Wan、C. Sun、C. Gao、N. Sang 和 C. Li，“ 用于视觉跟踪的层次化特征嵌入 ”，载于欧洲计算机视觉会议，2022。[138] X. Chen、H. Peng、D. Wang、H. Lu 和 H. Hu，“SeqTrack：用于视觉目标跟踪的序列到序列学习”，载于 CVPR，2023，第 14 572–14 581 页。[139] X. Wei、Y. Bai、Y. Zheng、D. Shi 和 Y. Gong，“自回归视觉跟踪”，载于 CVPR，2023，第 9697–9706 页。[140] X. Chen、B. Yan、J. Zhu、D. Wang、X. Yang 和 H. Lu，“Transformer 跟踪”，载于 CVPR，2021，第 8126–8135 页。[141] F. Xie、C. Wang、G. Wang、Y. Cao、W. Yang 和 W. Zeng，“相关性感知深度跟踪”，载于 CVPR，2022。[142] L. Lin、H. Fan、Y. Xu 和 H. Ling，“SwinTrack：用于 Transformer 跟踪的简单而强大的基线”，载于神经信息处理系统大会，2022。[143] L. Hong、S. Yan、R. Zhang、W. Li、X. Zhou、P. Guo、K. Jiang、Y. Chen

，J. Li ，Z. Chen 等 ，“OneTracker：统一视觉目标跟踪与基础模型及高效调优， ”发表于 CVPR，2024年，第 19 079–19 091 页。[144] W. Kay、J. Carreira、K. Simonyan、B. Zhang、C. Hillier、S. Vijayanarasimhan、F. Viola、T. Green、T. Back、P. Natsev、M. Suleyman 和 A. Zisserman，《Kinetics 人类动作视频数据集》，发表于 arXiv:1705.06950，2017年。[145] J. Carreira、E. Noland、A. Banki-Horvath、C. Hillier 和 A. Zisserman，《关于 Kinetics-600 的简短说明》，发表于 arXiv:1808.01340，2018年。[146] J. Carreira、E. Noland、C. Hillier 和 A. Zisserman，《关于 Kinetics-700 人类动作数据集的简短说明》，发表于 arXiv:1907.06987，2019年。[147] M. Monfort、A. Andonian、B. Zhou、K. Ramakrishnan、S. A. Bargal、T. Yan 和 A. Oliva，《Moments in Time 数据集：用于事件理解的百万级视频》，发表于 IEEE模式分析与机器智能汇刊，第42卷，第2期，第502–508页，2019年。[148] M. Bain、A. Nagrani、G. Varol 和 A. Zisserman，《Frozen in Time：用于端到端检索的联合视频和图像编码器》，发表于 IEEE国际计算机视觉会议，2021年。[149] T. Y. Lin、M. Maire、S. Belongie、J. Hays、P. Perona、D. Ramanan 和 C. L. Zitnick，《Microsoft COCO：常见物体及其场景》，发表于 欧洲计算机视觉会议，2014。[150] M. Muller、A. Bibi 和 G. S，《TrackingNet：面向野外目标跟踪的大规模数据集与基准》，发表于欧洲计算机视觉会议，2018年，第300–317页。[151] N. Xu、L. Yang、Y. Fan、D. Yue、Y. Liang、J. Yang 和 T. Huang，《YouTube-VOS：大规模视频目标分割基准》，发表于 arXiv:1809.03327，2018年。[152] Y. Wu、J. Lim 和 M.-H. Yang，《目标跟踪基准》，发表于 IEEE模式分析与机器智能汇刊，第37卷，第9期，第1834–1848页，2015年。[153] X. Li、Q. Liu、W. Pei、Q. Shen、Y. Wang、H. Lu 和 M. Yang，《信息丰富的跟踪基准》，发表于 arXiv:2112 06467，2021年。.[154] M. Danelljan、G. Bhat、F. S. Khan 和 M. Felsberg，《ATOM：通过重叠最大化实现精准跟踪》，发表于 CVPR，2019年。[155] F. Perazzi、J. Pont-Tuset、B. McWilliams、L. V. Gool、M. Gross 和 A. Sorkine-Hornung，《视频目标分割的基准数据集与评估方法》，发表于 CVPR，2016。[156] Z. Wang、J. Xu、L. Liu、F. Zhu 和 L. Shao，《RANet：用于快速视频目标分割的排序注意力网络》，发表于 IEEE国际计算机视觉会议，2019年。[157] A. Robinson、F. Lawin、M. Danelljan、F. Khan 和 M. Felsberg，《学习如何学习视频目标分割的快速鲁棒目标模型》，发表于 CVPR，2020年。[158] Y. Zhang、Z. Wu、H. Peng 和 S. Lin，《一种用于视频目标分割的传导式方法》，发表于 CVPR，2020年。[159] G. Bhat、F. Lawin、M. Danelljan、A. Robinson、M. Felsberg、L. Gool 和 R. Timofte，《学习视频目标分割中该学习什么》，发表于 欧洲计算机视觉会议，2020年。[160] Z. Yang、Y. Wei 和 Y. Yang，《通过前景-背景整合实现协同视频目标分割》，发表于 欧洲计算机视觉会议，2020年。[161] B. Duk、A. Ahmed、C. Wolf、P. Aarabi 和 G. W. Taylor，《SSTVOS：用于视频目标分割的稀疏时空变换器》，发表于 CVPR，2021年。[162] M. Paul、M. Danelljan、C. Mayera, 和 L. V. Gool，《通过分割实现鲁棒视觉跟踪》，发表于 欧洲计算机视觉会议，2022年。[163] K. Maninis、S. Caelles、Y. Chen、J. Pont-Tuset、L. Leal-Taix´e、D. Cremers 和 L. Van Gool，《无需时间信息的视频目标分割》，发表于 IEEE模式分析与机器智能汇刊，2018年。[164] J. Cheng、Y. Tsai、W. Hung、S. Wang 和 M. Yang，《通过部件跟踪实现快速、准确的在线视频目标分割》，发表于 CVPR，2018年。[165] Y. Hu、J. Huang 和 A. Schwing，《 VideoMatch：基于匹配的视频目标分割》，发表于 欧洲计算机视觉会议，2018年。[166] Q. Wang、L. Zhang 和 L. Bertinetto，《快速在线目标跟踪与分割：一种统一方法》，发表于 CVPR，2019年。[167] A. Lukezic、J. Matas 和 M. Kristan，《D3S：一种判别式单次分割跟踪器

,” 在*CVPR* , 2020。

[168] B. Yan、Y. Jiang、P. Sun、D. Wang、Z. Yuan、P. Luo 和 H. Lu，《迈向目标跟踪的大统一》，载于欧洲计算机视觉会议，2022年。[169] N. Mayer、E. Ilg、P. Hausser、P. Fischer、D. Cremers、A. Doso- vitskiy 和 T. Brox，《用于训练卷积网络进行视差、光流和场景流估计的大型数据集》，载于 CVPR，2016年，第 4040–4048 页。[170] Z. Qi、R. Dong、G. Fan、Z. Ge、X. Zhang、K. Ma 和 L. Yi，《对比与重建：生成式预训练引导的对比式三维表征学习》，载于国际机器学习会议，2023年。[171] Y. Pang、W. Wang、F. E. Tay、W. Liu、Y. Tian 和 L. Yuan，《用于点云自监督学习的掩码自编码器》，载于欧洲计算机视觉会议，2022年，第 604–621 页。[172] H. Jhuang、J. Gall、S. Zuffi、C. Schmid 和 M. J. Black，《迈向动作识别理解》，载于 IEEE国际计算机视觉会议，2013年。[173] X. Wang、A. Jabri 和 A. A. Efros，《从时间的循环一致性中学习对应关系》，载于 CVPR，2019年。[174] X. Li、S. Liu、S. De Mello、X. Wang、J. Kautz 和 M.-H. Yang，《用于时间对应关系的联合任务自监督学习》，神经信息处理系统大会，第 32 卷，2019年。[175] X. Chen 和 K. He，《探索简易孪生表征学习》，载于 CVPR，2021年，第 15 750–15 758 页。[176] D. Gordon、K. Ehsani、D. Fox 和 A. Farhadi，《看着世界流逝：从未标注视频中进行表征学习》，arXiv:2003.07990，2020年。[177] S. Purushwalkam 和 A. Gupta，《揭秘对比式自监督学习：不变性、数据增强与数据集偏差》，神经信息处理系统大会，第 33 卷，第 3407–3418 页，2020年。[178] R. Girdhar、A. El-Nouby、M. Singh、K. V. Alwala、A. Joulin 和 I. Misra，《OMNIMAE：在图像和视频上进行单模型掩码预训练》，载于 CVPR，2023年，第 10 406–10 417 页。[179] X. Sun、P. Chen、L. Chen、T. Li、M. Tan 和 C. Gan，《M3 视频：用于自监督视频表征学习的掩码运动建模》，载于 CVPR，2023年。[180] T. Yao、Y. Zhang、Z. Qiu、Y. Pan 和 T. Mei，《SeCo：探索用于无监督表征学习的序列监督》，载于 AAAI，第 35 卷，第 12 期，2021年，第 10 656–10 664 页。[181] Z. Yin、T. Darrell 和 F. Yu，《用于匹配密度估计的层次离散分布分解》，载于 CVPR，2019年。[182] S. Jiang、D. Campbell、Y. Lu、H. Li 和 R. Hartley，《利用全局运动聚合学习估计隐藏运动》，载于 IEEE国际计算机视觉会议，2021年，第 9772–9781 页。[183] Z. Zheng、N. Nie、Z. Ling、P. Xiong、J. Liu、H. Wang 和 J. Li，《DIP：用于高分辨率光流的深度逆向 PatchMatch》，载于 CVPR，2022年，第 8925–8934 页。[184] D. Sun、C. Herrmann、F. Reda、M. Rubinstein、D. J. Fleet 和 W. T. Freeman，《解耦光流的架构与训练》，载于欧洲计算机视觉会议，2022年，第 165–182 页。[185] Y. Lu、Q. Wang、S. Ma、T. Geng、Y. V. Chen、H. Chen 和 D. Liu，《TransFlow：作为光流学习器的 Transformer》，载于 CVPR，2023年，第 18 063–18 073 页。[186] C. Deng、A. Luo、H. Huang、S. Ma、J. Liu 和 S. Liu，《用于高效光流估计的显式运动解耦》，载于 IEEE国际计算机视觉会议，2023年，第 9521–9530 页。[187] H. Morimitsu、X. Zhu、X. Ji 和 X.-C. Yin，《用于高效光流估计的循环部分核网络》，载于 AAAI，第 38 卷，第 5 期，2024年，第 4278–4286 页。[188] N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhutdinov，《Dropout：一种防止神经网络过拟合的简单方法》，载于机器学习研究期刊，2014年。