视频分析的关键技术

- 视频动作识别
  - 传统方法：特征分析
  - 双流卷积网络：分别建模空间（外观）和时间（运动）信息
  - FlowFace：FlowFace提出了一种两阶段流水线，结合创新的2D对齐网络和3D模型拟合优化，显著提升跟踪精度
    - 密集2D对齐网络
      - 构架设计
        - 图像特征编码：采用Segformer（基于Vision Transformer）提取多尺度图像特征。
        - UV位置编码模块：通过多尺度纹理金字塔生成UV空间的位置嵌入，增强结构一致性。
        - 改进的RAFT模块：将光流预测网络RAFT适配为UV到图像的密集流（UV-image flow），并增加不确定性预测。
      - 训练目标
        - 使用高斯负对数似然（GNLL）损失监督顶点位置和密集流的预测，通过迭代优化逐步细化结果。
        - 引入随机背景替换、遮挡增强等数据增强策略，提升模型鲁棒性。
    - 3D模型拟合
      - 参数化模型选择：基于FLAME模型（5023顶点），引入身份参数、表情参数、骨骼姿态和静态顶点形变
      - 能量函数优化
        - 对齐能量：最小化预测顶点投影与2D对齐的误差。
        - 正则化项：约束FLAME参数、时间平滑性（加速度约束）、MICA中性形状先验（提升身份-表情解耦）和顶点形变惩罚。
        - 多视图联合优化：支持多相机输入，通过共享身份参数和优化相机外参提升重建一致性。
- 视频生成与预测
  - 生成对抗网络（GAN）：通过分离前景和背景动态，实现了逼真的视频生成
  - 变分自编码器（VAE）：一种基于自编码器结构的深度生成模型，结合了概率建模和变分推断的思想，主要用于数据生成、特征学习以及潜在空间的连续采样。
  - 多尺度视频预训练（MVP）：通过多尺度上下文预测学习视频的时空结构
    - 关键设计
      - 多尺度时间聚合：将视频分割为8帧的非重叠片段，构建观测序列和未来序列，通过不同时间跨度聚合未来片段的上下文信息。
      - 时空自注意力机制：采用多尺度视觉变换器（MViT）作为基础编码器，结合多头自注意力（MHA）捕捉时空区域间的动态关系，避免传统全局表示丢失细粒度信息。
      - 对比损失优化：通过对比学习最大化预测的未来区域表示与真实聚合表示之间的相似性，同时抑制负样本（其他视频或不同时空区域的表示）
    - 与现有方法区别
      - CVRL/CPC：仅关注片段级相似性，而MVP强调对未来上下文的因果推理。
      - LSTCL/CONSTCL：虽涉及长短期对比，但未显式建模多尺度动态，MVP通过多步预测和随机时间偏移提升泛化能力。
- 多模态学习：旨在融合视频、音频和文本等多种模态信息。
- 自监督学习：在于利用数据本身的结构或特性自动生成监督信号，而非依赖外部人工标注的标签