ChatGPT

当前主流与最新视频检索技术调研

文本描述检索视频片段（Text-to-Video Retrieval）

- **任务定义与挑战**：文本到视频检索旨在根据自然语言描述查询从视频库中检索相关视频片段，主要挑战在于跨模态语义对齐：需要将文本和视频映射到同一语义空间，度量其相似性。常见做法包括基于**对比学习**的全局对齐（如CLIP风格的双编码器）和基于**Transformer跨模态编码器**的交互对齐等。比如，**CLIP4Clip**将预训练的图像-文本模型CLIP知识迁移到视频检索，通过端到端学习获得了MSR-VTT、ActivityNet等数据集的SOTA效果 [1]；**Frozen-in-Time**使用时空Transformer对视频进行编码，采用“图像视作冻结帧”再逐步学习时序上下文的预训练策略，在MSR-VTT、MSVD、DiDeMo、LSMDC等基准上取得了顶尖成绩 [2] [3]。
- **代表模型**：除了CLIP4Clip和Frozen-in-Time外，还出现多种预训练或专用模型：例如对比预训练的**VideoCLIP**（Xu等，2021）通过联合训练视频和文本对（无标签）取得了序列级检索和视频问答任务的SOTA性能 [4]；**VIOLET**（Fu等，2021）设计了端到端视频-语言Transformer并提出了Mask Visual-token Modeling任务，同样在多项视频问答和文本检索任务上刷新了记录 [5]；Salesforce提出的**ALPRO**（Align & Prompt）引入视频-文本对比损失和实体提示任务，显著提升了视频语言预训练效果 [6]。另有使用对比损失或单纯基于CLIP的增强方法，如**CLIP2Video**、**CLIP-ViP**、**BLIP**（零样本文本检索上表现尤为强劲 [7]）等。
- **编码器与技术路线**：文本编码器多采用Transformer（如BERT或GPT），视频编码器包括3D卷积网络（I3D、SlowFast、TSM等）或时空Transformer（TimeSformer、VideoSwin等）。两者通常通过**双塔结构+对比损失**对齐（代表双编码器方法），或在初步对齐后再接多模态交叉编码器（代表联合编码方法）来优化匹配。部分工作还利用对象检测、语义标签或字幕信息进行辅助对齐。
- **训练数据集**：常用的下游检索数据集包括**MSR-VTT**、**MSVD**、**DiDeMo**、**LSMDC**、**ActivityNet Captions**、**YouCook2**等，这些数据集提供视频-文本对以评测检索性能 [8]。预训练时常用**WebVid-2M**、**HowTo100M**、**CC3M/CC12M**等大规模视频或图文对数据，以增强模型表征能力。
- **评估指标**：通常使用**Recall@K**（检索结果中前K位包含相关视频的比例）和**Median Rank**等指标来评测检索质量 [8] [9]。Recall@K数值越高（如R@1/R@5/R@10），模型性能越佳；中位排名越低（MedR）说明排名更靠前。

视频中的人物识别（Person Identification in Videos）

- **跨模态人物检索**：除了文本查询外，**图像检索视频人物**通常对应视频行人重识别（Video Person Re-ID）任务，即给定一张目标人物图像，从视频序列中检索出相同身份的行人序列。这类方法通常先使用行人检测器（如YOLO、Faster R-CNN）在视频帧中检测人物，再提取视频帧的人物特征并进行匹配。研究表明，视频重识别利用时序信息提取更丰富的特征，可以显著超过静态图像检索 [10] [11]。典型方法结合了CNN+RNN/Transformer等结构来编码时空特征，并使用度量学习损失优化不同摄像头下的身份匹配。
- **文本描述人物检索**：最近有工作专门面向**文本检索视频中人物**。如TVPR（Text-to-Video Person Retrieval）任务构建了跨模态视频人物数据集TVPReid，视频中每个行人视频配有详细自然语言描述 [12]。该任务使用多模态学习方法，将文本和整段视频编码对齐，从而应对遮挡、动作变化等问题 [13] [14]。TVPReid上的MFGF模型（多元素特征引导分片学习）通过逐步学习文本-视觉和文本-动作的对齐信息，在新任务上达到了最先进性能 [12] [15]。
- **视频人脸识别与行为识别**：视频分析还包括对显著人物的**人脸识别**，即在视频帧中检测人脸并识别身份，通常先用人脸检测器定位，再用预训练深度人脸模型（如FaceNet、ArcFace等）提取特征并匹配。此技术常用于视频中给人物打标签（人物标签学习）。此外，**行动识别**（行为识别）关注视频中人的动

作分类（如行走、奔跑、打篮球等），模型如C3D、I3D、SlowFast、Timesformer等通过学习时空动态来识别视频中的人类活动，这也是视频人物分析的重要组成部分。

- **Zero-shot / Few-shot 识别**：为了处理新出现的身份或缺少标注的问题，研究者还探索了**零样本/少样本人物识别**。例如针对少量训练样本的视频行人重识别，吴林等人提出了基于对抗学习的少样本方法，以学习具有视角不变性的时序特征 [16]。类似地，也有基于属性描述或生成模型的零样本人脸识别方法，用于扩展模型到未见过的人物。

带图像+文本查询的视频检索（Image+Text Query Video Retrieval）

- **任务定义**：多模态联合查询（图像+文本）的视频检索任务即“组合检索”。一个典型场景称为**组合视频检索**（Composed Video Retrieval, CoVR）：给定一个参考视频或图像和一段**修改说明文本**（描述查询意图的差异），检索数据库中符合该描述的目标视频 [17] [18]。例如查询：“人穿红色上衣的篮球比赛视频”，其中图像提供初始场景，文本指定“红色上衣”。类似地，组合图像检索（CoIR）是图像版的组合查询。该任务在可视化搜索、电子商务、体育赛事检索等场景有实际意义 [18]。
- **代表模型与技术框架**：目前方法通常将图像特征和文本特征分别编码，然后采用融合策略匹配视频特征。早期基线CoVR-BLIP [19] 使用BLIP模型：先用视觉编码器提取查询图像特征，再结合修改文本通过“图像引导文本编码器”生成查询向量，与视频编码对比学习对齐。最新方法（如Thawakar et al. 2024）进一步引入详细的语义描述和多模态对齐策略 [20] [21]，同时训练视觉、文本和多模态嵌入。实验表明，对比CoVR-BLIP，新方法在WebVid-CoVR等数据集的Recall@1可提高约7% [22] [19]。融合方式包括简单平均、全连接投影和跨模态注意力等，不同工作会选用最优的融合结构。
- **典型数据集**：为支持CoVR任务，**WebVid-CoVR**数据集被构造出来，其中训练集由131K个“源视频+修改文本+目标视频”三元组（Synthetic生成）组成，测试集则为人工筛选的高质量样本 [23]。修改文本平均长度约4.8词，测试集含3.2K条三元组 [23]。CoIR任务常用数据集如FashionIQ等，用于验证模型的零样本迁移能力。
- **应用场景**：多模态检索可用于精细化检索需求，如**电商与时尚**场景中根据一张衣服图片加属性（“红色”、“长袖”）检索类似服装；在**体育赛事**中查找指定球员的比赛片段；在**监控视频**中根据示例图像和补充描述定位特定人物等 [18]。通过融合图像和文本信息，系统能更准确理解用户意图、排除歧义，从而实现更精准的检索。

参考文献：以上内容基于近期学术文献与技术报告归纳总结 [2] [1] [7] [4] [5] [8] [9] [12] [10] [20] [23]。

---

1 [2104.08860] CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval
https://arxiv.org/abs/2104.08860

2 3 Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval
https://openaccess.thecvf.com/content/ICCV2021/papers/
Bain_Frozen_in_Time_A_Joint_Video_and_Image_Encoder_for_ICCV_2021_paper.pdf

4 [2109.14084] VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding
https://arxiv.org/abs/2109.14084

5 [2111.12681] VIOLET : End-to-End Video-Language Transformers with Masked Visual-token Modeling
https://arxiv.org/abs/2111.12681

6 [2112.09583] Align and Prompt: Video-and-Language Pre-training with Entity Prompts
https://arxiv.org/abs/2112.09583

7 BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation
https://arxiv.org/pdf/2201.12086

8 EA-VTR: Event-Aware Video-Text Retrieval
https://arxiv.org/html/2407.07478v1

9 Leveraging Auxiliary Information in Text-to-Video Retrieval: A Review
https://arxiv.org/html/2505.23952v1

10 11 Deep Learning for Video-based Person Re-Identification: A Survey
https://arxiv.org/html/2303.11332v3

12 13 14 15 TVPR: Text-to-Video Person Retrieval and a New Benchmark
https://arxiv.org/html/2307.07184v3

16 [1903.12395] Few-Shot Deep Adversarial Learning for Video-based Person Re-identification
https://arxiv.org/abs/1903.12395

17 18 19 20 21 22 23 Composed Video Retrieval via Enriched Context and Discriminative Embeddings
https://openaccess.thecvf.com/content/CVPR2024/papers/
Thawakar_Composed_Video_Retrieval_via_Enriched_Context_and_Discriminative_Embeddings_CVPR_2024_paper.pdf