近年来，互联网上的多媒体数据迅速增加，提供了更丰富和全面的信息。然而，从这些大规模多模态内容中找到真正需要的信息，成为一个新的挑战。跨模态信息检索正是为了解决这一问题，通过用户输入任意媒体类型的数据，检索出所有媒体类型中的语义相关数据。这一任务具有重要的应用价值，如从已知目标文字检索相似的图像，汇聚不同媒体数据展示多视角舆情事件演化过程，以及基于多媒体内容关联不同社交网络账户等。

跨模态信息检索主要面临以下两个技术挑战：

- **异构鸿沟**：不同模态的信息（如视觉和语言）的数据分布和特征表示不一致，如何实现多模态信息的语义关联和模态跨越是首要挑战。
- **语义鸿沟**：计算机特征表示与人类语义概念不一致，如何利用多模态和细粒度信息缩短语义鸿沟也是关键问题。

跨模态信息检索事实上解决的是特征向量的构建问题，具体到检索算法，在大数据处理领域有大量算法优化检索性能：

- **基于相似性度量的Brute force**：遍历所有数据点进行相似性计算，复杂度较高，但可以保证结果的准确性。
- **近似最近邻（ANN）**：通过构建数据索引结构，如KD树、LSH等，加速相似性计算，降低时间复杂度。
- **局部敏感哈希（LSH）**：将高维数据投影到低维空间，利用哈希函数加速相似性检索，适用于大规模数据集。

跨模态信息检索的粗粒度检索

粗粒度跨模态检索主要包括图像文本检索和视频文本检索，即给定文本或视觉查询，从另一个模态的数据库中找到语义最相似的样本。该方法侧重整体语义相似度，忽略数据之间的细节差异。

粗粒度跨模态检索模型架构主要包括以下几个部分：

1. **视觉特征提取**：使用卷积神经网络（CNN）或视觉Transformer从图像或视频中提取全局特征。
    - 传统的图像特征提取方法包括局部纹理特征（LBP）、尺度不变特征（SIFT）、局部梯度特征（HOG）等，这些方法存在严重的“语义鸿沟”。深度学习方法（如AlexNet、VGG、ResNet等）在各类视觉任务中展示了有效的语义表征能力。
    - 视频特征提取包括2D-CNN和3D-CNN，前者在多个帧上进行2D卷积，没有考虑到视频的时域信息，后者能够同时捕获视频中的空域与时域信息，表示静态视觉对象和动态视觉事件的特征。
2. **文本特征提取**：使用循环神经网络（RNN）、长短期记忆网络（LSTM）或Transformer从文本中提取特征。
3. **对齐模块**：将图像和文本特征映射到公共语义表示空间中，以消除不同模态之间的模态鸿沟，从而度量视觉和文本的相似度。

这些方法通过不同的损失函数（如排名损失、对比损失等）和训练策略，提升了跨模态检索的性能。

论文发展关系：

- **Use What You Have: Video Retrieval Using Representations From Collaborative Experts. BMVC 2019** 提出了利用多个专家模型并行对视频进行表示学习的方法，称为Collaborative Experts。该方法通过门控机制进行多维度信息融合，利用多个专家模型（如图像对象、场景识别、行为识别等）对视频进行多方面的理解和表征，从而提升了视频检索的性能。具体而言，每个专家模型都专注于不同的视觉特征，如颜色、纹理、形状等，通过综合这些特征，模型能够更全面地理解视频内容，提高了检索的准确性。

- **Stacked Cross Attention for Image-Text Matching. ECCV 2018** 在 BMVC 2019 的基础上，进一步引入了交叉注意力机制，以提升图像和文本匹配的精度。该方法在图像和文本的匹配过程中，利用交叉注意力机制来捕获细粒度的对齐信息，从而提升了检索效果。具体实现上，模型通过在图像和文本之间多次迭代注意力机制，逐步细化匹配特征，提高了语义对齐的准确性。
- **Graph Structured Network for Image-Text Matching. CVPR 2020** 提出了一种基于图结构的网络模型GSMN（Graph Structured Network），通过构建图结构，将目标、关系和属性明确建模为短语，并通过对短语对应关系的学习进行细粒度对齐关系的学习，从而提升图文匹配的效果。具体来说，GSMN模型利用图神经网络（GNN）对图结构中的节点和边进行信息传播和更新，实现了图像和文本之间更精细的语义对齐。
- **Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. AAAI 2020** 提出了 Unicoder-VL 模型，通过在大规模数据集上进行预训练，获取图像和文本的通用表征。这种方法不仅能够处理图像-文本匹配问题，还能够适用于其他跨模态任务，进一步提升了模型的泛化能力和性能。具体实现上，Unicoder-VL利用Transformer架构，通过自监督学习方法在不同模态的数据上进行预训练，从而学得了统一的表示空间，实现了跨模态任务的高效处理。
- **TALL: Temporal Activity Localization via Language Query. ICCV 2017** TALL模型通过建立时间回归模型，对预设的时间片段做偏移回归调整时间框的大小，解决了视频片段的定位问题。该模型通过视频片段编码、查询文本编码和跨模态特征融合，实现了文本和视频片段的对齐和检索。具体来说，TALL模型首先使用C3D网络对视频片段进行表征，然后利用LSTM对查询文本进行编码，最后通过跨模态注意力机制将视频和文本特征进行融合，从而实现了视频片段的准确定位。
- **Learning 2D Temporal Adjacent Networks for Moment Localization with Natural Language. AAAI 2020** 在TALL的基础上提出了一种直接定位的方法，通过计算每个时间点可能是目标起止时间点的概率，直接预测视频片段的起始点和终止点。这种方法消除了对视频的重复处理和计算，提高了检索效率。具体实现上，模型构建了2D时间图，将候选片段特征填入其中，并进行特征融合和卷积，从而获取全局特征，进行回归训练。
- **Localizing Natural Language in Videos. AAAI 2019** 2D-TAN模型利用2D时间图建模不同时刻之间的时序关系，使每段片段都具有上下文信息，从而提高了视频片段定位的精度。该方法通过构建2D时间图，将候选片段特征填入其中，并进行特征融合和卷积，获取全局特征，进行回归训练。

跨模态信息检索的细粒度检索

细粒度跨模态检索任务侧重于捕捉数据之间的细微差异，通常需要更加复杂的模型来捕捉局部语义信息进行对齐，如服饰检索、行人检索等。

论文发展关系：

- **LapsCore: Language-Guided Person Search via Color Reasoning. ICCV 2021** 通过设置文本引导图像着色与图像引导文本填词两个颜色推理子任务，引导模型学习细粒度跨模态关联关系。这种方法通过颜色的理解和推理，实现了更精确的文本到图像的行人检索。具体来说，模型通过对灰度图像进行着色任务，迫使其理解文本描述中的颜色信息，同时，通过图像引导文本填词任务，使模型在文本理解上更准确地捕捉到颜色和对象之间的关联。
- **Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. CVPR 2023** 在 ICCV 2021 的基础上，提出了隐式关系推理和对齐方法，基于CLIP模型，通过多模态交互编码器学习细粒度对齐信息，提升了文本到图像行人检索的精度。该方法在测试阶段可以去除交互模块，只利用全局特征进行检索，保证了检索的效率和精度。具体实现上，模型通过在CLIP的基础上增加隐式关系推理层，实现了更细粒度的文本与图像对齐，从而提升了检索性能。

- **RaSa: Relation and Sensitivity Aware Representation Learning for Text-Based Person Search. IJCAI 2023** 基于ALBEF架构，通过动量编码器生成困难负样本，进一步提升了模型的细粒度对齐能力。动量编码器生成的困难负样本，使得模型能够更好地学习到文本和图像之间的细粒度关联关系，从而提升了文本行人检索的性能。具体来说，动量编码器通过缓慢更新参数，生成了更具挑战性的负样本，促使模型在训练过程中更加准确地学习到文本和图像之间的细粒度对齐信息。

跨模态信息检索的交互式检索

交互式跨模态检索通过与用户的交互，逐步细化检索需求和结果，如单轮组合检索、多轮对话图文检索等。

论文发展关系：

- **Composing Text and Image for Image Retrieval-An Empirical Odyssey. CVPR 2019** 提出了一种组合图像检索的方法，通过结合图像和文本描述，提升图像检索的灵活性和精度。具体来说，模型通过简单的特征组合，初步实现了图像和文本的组合检索，并验证了这种方法在多种检索任务中的有效性。
- **Context-I2W: Mapping Images to Context-Dependent Words for Accurate Zero-Shot Composed Image Retrieval. AAAI 2024** 提出了Context-I2W模型，通过上下文依赖的词映射方法，将图像中的视觉内容映射到伪单词标记，并在推理过程中，将图像映射到伪单词标记，形成统一的语言空间查询。具体来说，模型通过上下文选择器和视觉目标提取器，更加精确地捕捉图像和文本之间的对应关系，实现了更高精度的组合图像检索。
- **Chatting Makes Perfect: Chat-Based Image Retrieval. NIPS 2024** 提出了对话式图文检索，通过与用户的多轮对话，逐步获取用户的检索需求。模型不仅通过图像搜索获取符合需求的图像，还通过对话构建生成下一轮对用户的问题，以获得更多的附加信息，实现更加自然和渐进式的检索。具体实现上，模型通过对用户输入的解析和理解，逐步引导用户提供更详细的描述，从而实现了更高效和精准的检索。