E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391;TP18　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3144-12

论文引用格式:Hu Y M, Wang X H, Deng B, Zhao Y Y and Zhao Y Y. 2026. A survey on multimodal information fusion. Journal of Image and Graphics, 31(9):3144-3155(胡玉梅, 王晓华, 邓豹, 赵洋洋, 赵艺阳. 2026. 多模态信息融合研究进展. 中国图象图形学报, 31(9):3144-3155)[DOI:10.11834/jig.260236]

多模态信息融合研究进展

胡玉梅*,王晓华,邓豹,赵洋洋,赵艺阳

中国航空工业集团西安航空计算技术研究所,西安　710065

摘　要:随着传感器网络、信息感知方式及大数据处理技术的飞速发展,面对来自于不同类型、不同特性和不同感知方式的信息,及其呈现出的多域、多模态、模糊化以及数据关联不完备特点,多模态信息融合技术以其类人式自动感知与强大的综合推理能力的优势受到越来越多的关注。本文首先介绍了多模态信息融合的意义和必要性,进而,在融合架构层面梳理了从早期融合、晚期融合到混合融合的演进脉络,揭示了各架构在信息保留与计算效率间的权衡关系。同时,详细阐述了基于注意力的交互建模、基于对比学习的语义对齐以及以大语言模型为载体的生成式融合3种前沿多模态融合方式。继而,介绍了COCO(common objects in context)、LAION-400M(large-scale artificial intelligence open network-400M)、Visual Genome、MS COCO Captions、Conceptual Captions、Flickr30k数据集和航空相关数据集等典型多模态数据集及相应应用领域。并且,进一步探讨了多模态融合在包括战场态势感知、多源情报分析与无人系统协同的典型军事领域应用。此外,指出随着对比学习和多模态预训练的成熟,高质量的单模态表示已不再是瓶颈,研究焦点由早期如何将异质模态映射到统一空间,转向于如何在表示对齐的基础上设计能够捕捉模态间复杂、动态,甚至矛盾关系的交互机制,并基于融合架构、融合模型和计算成本给出多模态信息融合的3个发展方向。

关键词:多模态信息融合;多模态大语言模型;注意力机制;对比学习;生成式融合

A survey on multimodal information fusion

Hu Yumei*, Wang Xiaohua, Deng Bao, Zhao Yangyang, Zhao Yiyang

Xi'an Aeronautics Computing Technique Research Institute, AVIC, Xi'an　710065, China

Abstract: The rational and effective use of multisource information tend to expand the spatial and temporal coverage of measurements, fully excavate the target feature information contained in various sensors, greatly reduce the ambiguity of information, and improve detection performance. Faced with multimodal target data from different types, characteristics, and perception methods in some applications, such as in airborne multisensor target-tracking systems, various types of sensors with different working attributes (e.g., visible light images, infrared images, optical sensors, microwave radar, and LiDAR) are often involved, presenting measurement information of the target from different perspectives. The point cloud information obtained by LiDAR not only achieve high-precision ranging but also provide more accurate spatial information due to the varying reflectivities of different objects. However, this is susceptible to noise generated by complex environments and lacks semantic information. Compared to point cloud information, target image information provides rich surface textures and contextual semantic information that can more accurately restore the appearance and structure of objects; how-

收稿日期:2026-04-24;**修回日期**:2026-05-29;**预印本日期**:2026-06-05

*通信作者:胡玉梅　hym_henu@163.com

基金项目:航空科学基金项目(2024M066031002)

Supported by: Aeronautical Science Foundation(2024M066031002)

ever, it often lacks in-depth information. The current landscape is marked by the rapid development of sensor networks, information perception methods, and big data processing technologies. Faced with information perception from different types, characteristics, and methods, multimodal information fusion technology has received increasing attention. This is due to its advantages of human-like automatic perception and powerful comprehensive reasoning capabilities; its multi-domain and multimodal nature, as well as its ambiguity, and incomplete data association. The paper first introduces the significance and necessity of multimodal information fusion. Furthermore, at the fusion architecture level, it outlines the evolution from early fusion, late fusion, to hybrid fusion, revealing the trade-off relationship between information retention and computational efficiency in each architecture. At the same time, this paper elaborates on three cutting-edge multimodal fusion methods: attention-based interaction modeling, semantic alignment based on contrastive learning, and generative fusion with LLMs as the carrier. Then, it introduces typical multimodal datasets, such as the COCO dataset, LAION-400M dataset, Visual Genome dataset, MS COCO Captions dataset, Conceptual Captions dataset, Flickr30k dataset, and aviation-related datasets, along with their corresponding application fields. Moreover, the paper further explores the typical military applications of multimodal fusion, including battlefield situation awareness, multisource intelligence analysis, and unmanned system collaboration. This work also emphasizes that high-quality single-modal representations are no longer a bottleneck with the maturity of contrastive learning and multimodal pre-training. To date, the research focus has shifted from how to map heterogeneous modalities to a unified space in the early stage to how to design interaction mechanisms that can capture complex, dynamic, and even contradictory relationships between modalities based on representation alignment. On the basis of fusion architecture, fusion models, and computational costs, we propose three development directions for multimodal information fusion. We argue that the choice of multimodal fusion architecture directly impacts the degree of information retention, computational efficiency, and model interpretability. Thus, existing methods can be categorized into early fusion, late fusion, hybrid fusion based on the stage at which information fusion occurs. Early fusion, also known as data-level fusion, involves the integration of multisource information prior to or at the shallowest level of modality-specific feature extraction. This process preserves the richest original information and multimodal interaction details. Typical implementations include multimodal concatenation and multi-view encoding. Taking the visual-language model as an example, early fusion concatenates image patch embeddings and text word embeddings into a unified sequence, which is then fed into a joint encoder for processing. The advantage of this architecture lies in its ability to establish multimodal associations at a shallow level, thus facilitating the capture of fine-grained modality interactions. However, early fusion faces a severe dimensionality catastrophe problem. When the number of modalities increases or the feature dimensions of each modality become extremely high, the joint representation space grows exponentially. At the same time, during early fusion, the noise of some modalities may be amplified due to the varying noise characteristics of different modalities, resulting in poor representation quality. Meanwhile, late fusion adopts a diametrically opposite strategy in which each modality independently performs feature extraction and task prediction, with integration, weighted average, and meta-learner only at the final decision-making layer. In this "divide and conquer" design, the modal branches can be trained in parallel, achieving high computational efficiency; meanwhile, overfitting or noise in one modality is less likely to affect other modalities. Additionally, in scenarios wherein some modalities are missing, the remaining branches can still work normally, demonstrating strong system robustness. However, late fusion may lead to a higher consumption of computational resources as it requires training independent models for each modality. At the same time, independent models for each modality struggle to capture low-level interactions between them, making it difficult to model simple fusion at the decision-making level. Furthermore, data from different modalities may face alignment issues in time or space, such as the synchronization of video frames and audio signals. Hybrid fusion introduces cross-modal interaction in the middle layers of the network while maintaining modality-specific processing paths, aiming to combine the strengths of both. Its typical application is to use modality-independent encoders at the bottom layer, introduce a cross-attention module in the middle layer to achieve feature-level interaction, and separate again at the top layer to preserve modality-specific information. A deeper evolutionary direction is dynamic fusion, which allows the network to autonomously decide where and with what intensity information from various modalities can be fused. For example, a gated mechanism-based visual image and LiDAR multimodal fusion network dynamically adjusts the weights of information from each modality based on input data. Specifically,

it places more trust in visual images under strong lighting conditions, while giving higher weight to LiDAR point cloud information under low-light conditions. The choice of fusion architecture does not have an "optimal solution"; rather, it depends on the function of task characteristics and resource constraints. Generally, when there are fine-grained, location-related interactions between modalities (e. g. , image-text alignment), early fusion is better. When each modality can complete predictions independently and some modalities are prone to missing, late fusion is more robust. Meanwhile, dynamic fusion strikes a good balance between performance and robustness. Looking back at the development of multimodal fusion, the bottleneck of fusion has shifted from "representation" to "interaction". With the maturity of contrastive learning and multimodal pre-training, high-quality single-modality representation is no longer a bottleneck. In particular, the research focus has shifted from how to map heterogeneous modalities into a unified space in the early stage to how to design interaction mechanisms that can capture complex, dynamic, and even contradictory relationships between modalities based on representation alignment. Furthermore, fusion architecture has evolved from "static design" to "dynamic adaptation". In the real world, the correlation and reliability of modalities change dynamically with the environment, and the limitations of fixed fusion strategies are becoming increasingly evident. In military confrontations, electronic jamming may render specific sensors inoperative. Therefore, to enhance the robustness of the fusion system the activation state of modalities and fusion weights must be dynamically adjusted on the basis of input content and task context. The causal fusion model holds promise for breaking through the current limitations of relational learning. At present, most existing methods focus on learning statistical correlations between modalities, rather than causal relationships. This leads to two issues. First, the model is prone to learning spurious correlations; second, it struggles to generalize to environments beyond the training distribution. Therefore, introducing causal inference tools can enhance the fusion model's adversarial robustness and environmental transferability. Furthermore, a new multimodal fusion mechanism is designed in this paper. To date, redundant information from each modality may cause computational waste, and the corresponding noise information may affect fusion performance. Thus, how to utilize information theory techniques to identify and retain complementary information while suppressing redundancy and noise is a new development direction for multimodal fusion technology.

Key words: multimodal information fusion; multimodal large language model; attention mechanism; contrastive learning; generative fusion

0 引 言

多源信息的合理有效利用将扩展量测的空间和时间覆盖范围，充分挖掘各传感器包含的目标特征信息，大幅降低信息的模糊度，改进探测性能（何友等，2013；潘泉 等，2019）。面对来自于不同类型、不同特性以及不同方式感知的多模态目标数据（Li等，2025；Hangloo 和 Arora，2025），例如，在机载多传感器目标跟踪系统中，常涉及到可见光图像、红外图像、光学传感器、微波雷达和激光雷达等不同工作属性的多种类型传感器，从不同角度呈现目标的量测信息。激光雷达获取的点云信息不仅能够获取高精度测距，由于对不同的物体具有不同的反射率，还能提供更加准确的空间信息，但是有易受复杂环境影响产生噪点和缺少语义信息的缺陷。相对点云信息，目标图像信息提供丰富的表面纹理和上下文语义信息更能准确还原物体的外观和结构，但缺少深度信息。

在多源同质信息融合方面，融合理论和技术研究比较丰富并成功应用于战略预警与防御、多目标跟踪与识别和精确制导武器等军事领域，并逐渐辐射到智能交通、遥感监测、医学诊断、电子商务、人工智能、无线通信和工业过程监控与故障诊断等众多民用领域（Zhang 等，2025；Jiao 等，2024；Vakil 等，2021）。潘泉等人（2003，2012）梳理了同质信息融合系统建模方法、信息融合模型与系统设计以及未来发展方向，并系统性阐述现阶段存在的问题及其解决思路。潘泉等人（2019）针对基于联合优化的信息融合在目标跟踪领域的发展进行综述。随着机器学习和深度学习的快速发展和计算能力的提升，基于数据驱动的信息融合备受关注。Tzikas 等人（2008）首次尝试采用随机森林回归的方法学习从状态到量测之间的映射关系，以实现径向距—角度量测下的目标跟踪问题。文中通过概率模型仿真产生大量的目标状态和量测，并分析在不同数量的状态和量测

情况下，每个森林具有不同数目树的情况下算法性能。在应用方面，Song 等人（2016）采用高斯过程和支持向量机（support vector machine，SVM）的方法学习目标运动模型预测弹道系数，结合动态方程进行一系列的预测迭代以实现高速弹道目标在量测时刻瞬间的状态预测。针对图像融合问题，传统融合方法为包括加权平均、主成分分析（principal components analysis，PCA）变换和稀疏表示等的空间域技术以及包括金字塔变换、小波变换和非下采样轮廓波变换等变换域技术。

随着传感器网络、信息感知方式及大数据处理技术的飞速发展，感知手段日益丰富，感知信息日益呈现出多模态、模糊化以及数据关联不完备等特点。通过多传感器的融合，能有效降低单一传感器在恶劣环境下的失效率，增强感知系统的可靠性。例如，通过将激光雷达的高精度测距和空间几何探测优势与图像的语义、纹理优势互补，能够实现对复杂环境的高精度、鲁棒性综合感知。因此，多源、多域和多模态信息融合（multi-modal information fusion，MMIF）技术受到越来越多的关注。

1　多模态融合架构

多模态融合架构的选择直接影响信息保留程度、计算效率与模型可解释性。根据信息融合发生的阶段，现有方法可划分为早期融合（early fusion）、晚期融合（late fusion）和混合融合（hybrid fusion）等。

1.1　早期融合

早期融合，亦称数据级融合，在模型特征提取之前或最浅层完成多源信息的融合。早期融合过程中保留最丰富的原始信息和多模态交互细节。典型实现包括多模态拼接（Erukude 等，2025）（将文本与图像特征向量直接连接）、多视角编码（Chudasama 等，2022）（如 Transformer 的多模态输入序列）。以视觉—语言模型为例，早期融合将图像块嵌入与文本词嵌入拼接为统一序列，送入联合编码器处理。这种架构的优势在于：浅层即可建立多模态关联，有利于捕捉细粒度的模态交互。

然而，早期融合面临严峻的维度灾难问题。当模态数量增加或各模态特征维度过高时，联合表示空间呈指数级增长。同时，由于不同模态的噪声特性各异，早期融合过程中部分模态的噪声可能被放大，导致表征质量较差。

1.2　晚期融合

晚期融合采取截然相反的策略：各模态独立完成特征提取与任务预测，仅在最终决策层进行整合（投票（Rajpurkar 等，2016）、加权平均（Li 等，2025）和元学习器（Ruder，2017））。在这种“先分后总”的设计中各模态分支可并行训练，计算效率高；同时，某一模态的过拟合或噪声不易影响其他模态；此外，在部分模态缺失场景下，剩余分支仍可正常工作，系统鲁棒性较强。

然而，由于晚期融合需为每个模态训练独立的模型，可能导致更高的计算资源消耗。同时，各模态独立模型难以捕捉各模态间低层次的交互，决策层简单的加权、投票或者拼接难以建模和挖掘不同传感器量测之间的关联关系。此外，不同模态的数据可能存在时间或空间上的对齐问题，例如视频帧与音频信号的同步对齐。

1.3　混合融合

混合融合在网络的中间层引入跨模态交互，同时保持模态特定的处理路径，以试图结合二者之长。其典型应用是底层采用模态独立编码器（类似晚期融合），中层引入交叉注意力模块实现特征级交互（类似早期融合），顶层再次分离以保留模态特异性信息（Baltrusaitis 等，2019）。

更深层的演进方向是动态融合（Xue 和 Marculescu，2023），即让网络自主决定在何处、以何种强度融合各模态信息。例如，基于门控机制的视觉图像与激光雷达多模态融合网络根据输入信息动态调整各模态信息权重，即在强光照条件下更信任视觉图像，在低照度条件下激光雷达点云信息的权重更高。

融合架构的选择不存在“最优解”，而是任务特性与资源约束的函数。一般而言，当模态间存在细粒度、位置相关的交互时（如图像—文本对齐），早期融合更优；当各模态可独立完成预测且部分模态易缺失时，晚期融合更稳健；动态融合则在性能与鲁棒性间取得较好平衡。表 1 给出 3 种融合架构的对比。

2　多模态信息融合理论

人工智能在多模态信息融合领域得到了广泛应用，特别是随着大模型的爆发式发展，大模型以其统一表征、跨模态对齐、端到端学习和上下文建模等特

表1　3种融合架构的对比

Table 1　Comparison of three fusion architectures

融合架构	融合阶段	优势	劣势
早期融合	模型输入、特征提取初期	捕捉底层相关性，参数少，效率高	对齐难度大，易丢失单模态特性
晚期融合	预测输出、决策阶段	模块化设计，灵活性极强，鲁棒性高	缺乏深层特征交互，依赖单模态准确率
混合融合	特征层与决策层多级组合	兼顾特征互补与决策优化，表达力强	架构较复杂，训练和调优成本高

点在文本、图像、音频和视频等多模态理解和推理等复杂任务中提供了强大的工具（唐霖峰 等，2023）。根据多模态融合机理不同，主要分为3类技术范式：注意力机制（attention mechanism）、对比学习（contrastive learning）和生成式融合（generative fusion）。

2.1　基于注意力机制的交互建模

注意力机制模型（Brauwers 和 Frasincar，2023）通过模拟人眼观察目标时倾向于观察画面中某些重要特征，而忽视画面中不重要的特征的过程，使深度学习在提取事物特征时更有针对性。注意力机制主要包括自注意力机制、交叉注意力机制和多头注意力机制。

自注意力（self-attention）机制（Vaswani 等，2017）的核心思想是将输入特征同时映射为查询（Query）$\boldsymbol{Q}$、键（Key）$\boldsymbol{K}$和值（Value）$\boldsymbol{V}$向量，并通过计算$\boldsymbol{Q}$与$\boldsymbol{K}$之间的相似度来衡量不同位置的重要性。通过对$\boldsymbol{V}$进行加权求和，自注意力机制能够动态整合全局上下文信息，使每个位置的特征表示同时融合来自其他位置的关键信息。其实现原理为

$$A\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=f\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

式中，A表示注意力，函数f表示softmax函数，d_k表示向量$\boldsymbol{Q}$和$\boldsymbol{K}$的维数，自注意力机制架构如图1所示。

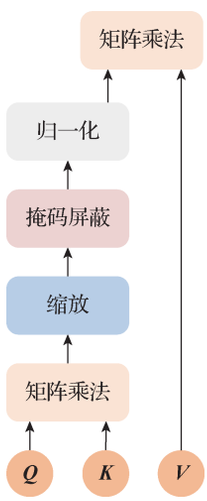


图1　自注意力机制架构

Fig. 1　The architecture of self-attention mechanism

由于在融合过程中模型动态聚焦于各模态中最相关的信息片段，而过滤不重要的信息，导致自注意力机制的有效信息抓取能力较弱，例如，自注意力机制无法利用图像的尺度信息和平移不变性以及图像的特征局部性。因此自注意力机制只有在大数据的基础上才能建立准确的全局关系。随着Transformer架构（Vaswani 等，2017；Devlin 等，2019）的提出，注意力机制由传统的局部特征增强逐步发展为具备全局依赖建模能力的核心模块。与基于卷积神经网络（convolutional neural network，CNN）的注意力机制（Wang等，2018）通过局部感受野进行表达不同，Transformer通过自注意力机制直接建模序列中任意位置之间的关系，为复杂场景下的全局语义理解提供了有效手段（Dosovitskiy 等，2021）。

交叉注意力（cross-attention）机制（Carion 等，2020）是自注意力的扩展，其核心是两个不同序列之间的注意力交互，从而使目标序列能够有选择地从源序列中获取有效信息，即查询（query）来自一个模态，键（key）和值（value）来自另一模态。例如，在视觉问答任务中，文本问题作为查询去“检索”图像中的相关区域，生成问题导向的视觉表示，实现“引导一个模态对另一模态的理解”的应用。

多头注意力（multi-head attention）机制（Vaswani 等，2017）不是一种独立的注意力类型，而是一种增强自注意力或交叉注意力性能的架构。实现原理表示为

$$\begin{aligned} MH\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)&=C\left(h_1,\cdots,h_h\right)\boldsymbol{W}^{O}\\ h_i&=A\left(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}\right)\end{aligned} \tag{2}$$

式中，MH表示多头注意力，C表示多头注意力融合函数，h_i表示第i个注意力头的注意力，$\boldsymbol{W}_i^{Q}$，$\boldsymbol{W}_i^{K}$，$\boldsymbol{W}_i^{V}$

和$\boldsymbol{W}^O$为参数矩阵。多头注意力架构如图2所示。

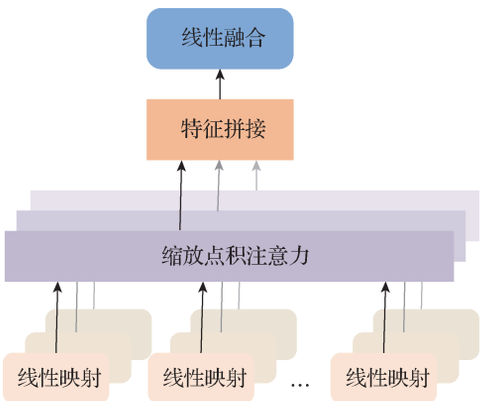


图2　多头注意力机制架构

Fig. 2　The architecture of multi-head attention mechanism

多头注意力机制是将多个自注意力或交叉注意力的“头”独立并行计算，然后将结果拼接，以增强模型的表达能力，适合并行化实现，从而提高计算效率。多头注意力进一步扩展了交互的多样性，其核心思想是将输入特征映射到多个不同的子空间，并在每个子空间中并行执行自注意力计算。多个注意力头可并行捕捉不同类型的跨模态关系，例如，不同注意力头分别关注全局语义对齐，聚焦局部细节匹配，建模时序依赖等。各注意力头的输出结果在通道维度上进行拼接，并通过线性变换进行融合。相比单一注意力结构，多头注意力机制能够更有效地捕获复杂的语义依赖关系。

Transformer架构正是建立在这一机制之上，通过堆叠多层交叉注意力实现深度的模态融合。Mushtaq等人(2025)在局部图特征基础上引入多头自注意力，首先将多路通道级特征拼接，然后进行$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$投影与注意力聚合，将离散的局部几何特征进行压缩以提升远距离稀疏目标的语义一致性，从而解决长距离依赖信息关联较差的问题。针对航拍图像中弱目标检测受杂乱背景、遮挡与小尺度影响导致鲁棒性不足的问题，Lan等人(2025)利用多头注意力特征融合模块首先提取全局特征与图像特征，然后学习多注意力的融合权重，以提升检测模型对弱小目标的适应能力。郭润泽等人(2025)设计一种基于信息感知引导的多头注意力融合检测网络模型，通过信息感知模块计算图像的光照信息和目标的局部对比度，引导多尺度差分注意力模块对可见光和红外图像的模态内和模态间特征进行深度交叉融合，以提升弱光条件下无人机对地目标的检测识别精度。同时基于多模吊舱、边缘计算模块和自组网电台构建了一套旋翼无人机多模目标检测系统。并在机载端对算法进行了轻量化部署和验证，结果表明在真实弱光场景下算法能显著提升无人机对目标的检测能力，且平均运行效率可达21.2帧/s。

注意力头的数量是一个需要权衡的超参数，从工程和架构设计的角度来看，往往是先根据经验确定一个合适的子空间大小(单头容量)，再反推得出头数。每个特征子空间需要足够大的维度(通常经验值为64或128)来承载复杂的语言特征。如果子空间太小则可能形成信息瓶颈，无法传递丰富的上下文信息。自注意力的时间和空间复杂度均为$O(N^2)$(N为序列长度)。当处理几十万甚至上百万词元的超长文本或高分辨率视频时，内存占用会呈爆炸式增长。为减小计算复杂度，学者们从稀疏注意力、线性注意力和IO (input/output)感知优化等方面开展相关研究。

值得注意的是，注意力机制的效能高度依赖模态表征的质量，即单模态编码器输出的特征本身。基于注意力机制的信息融合不是对弱特征的补救，而是对强表示的升华。表2给出各注意力机制特点的对比。

表2　各注意力机制特点

Table 2　Comparison of attention mechanisms

名称	自注意力机制	交叉注意力机制	多头注意力机制
核心关注对象	序列内部关系	序列之间关系	不同表示子空间
输入来源	单一序列(内部交互)	两个不同序列(跨序列交互)	多个并行的注意力头
主要功能	捕捉序列内部的长程依赖和上下文信息	实现不同序列或模态间的信息融合对齐	从多个子空间并行计算注意力，增强模型的表征能力
应用场景	编码器、语言建模	解码器、多模态模型	所有注意力机制的增强版本

2.2 基于对比学习的语义对齐

对比学习的兴起深刻重塑了跨模态融合学习范式：构造正样本对（语义匹配的跨模态数据）与负样本对（不匹配的数据），通过对比损失拉近正样本距离、推远负样本距离，从而在共享嵌入空间中实现模态对齐。

CLIP（contrastive language-image pre-training）（Radford等，2021）模型能够将图像和文本映射到共享语义空间，并计算生成文本的嵌入向量与目标区域视觉特征向量之间的匹配分数，量化语义描述的准确性，从而实现跨模态的匹配和理解，该模型被认为是基于语义对齐的里程碑式工作。通过对海量图文对进行对比预训练，CLIP模型中的对齐关系具有可泛化性，并使其具备强大的零样本迁移能力和开放词汇理解力，有效减少大量人工标注数据的依赖。得益于CLIP这类视觉—语言模型强大的零样本学习能力，Li等人（2024）提出一种基于外部视觉—名称记忆的语义生成方法，该方法构建了一个外部的视觉—名称记忆库，通过检索相关对象名称并将其与图像特征结合，生成更为准确的图像描述。然而，这类方法虽然能够感知图像整体内容，但是也会引入文本约束外的不相关噪声，导致生成语义质量较差。Su等人（2022）提出了一种基于预训练的GPT-2（generative pre-trained Transformer-2）和CLIP模型的零样本无训练语义生成方法，通过引入“魔幻分数”这一视觉控制项，计算图像与生成文本的相似性，从而指导文本生成，该方法相比于传统方法在解码生成速度上提高了27倍，大幅度提升了推理效率。

对比学习的优势在于：它无需昂贵的像素级标注（如图像分割掩码），仅需弱监督的配对数据即可学习强语义对齐。这使得模型可从互联网规模的弱标注数据中持续学习，突破了人工标注的成本瓶颈。然而图像与文本的对应关系往往是一对多的（同一图像可有多种有效描述），对比学习强制将其压缩为点对点的匹配可能导致语义多样性丢失。同时，由于对比学习难以显式建模模态间的因果关系（如“文本描述导致图像生成”与“图像内容引发文本描述”是不同的关系），因此在因果推理任务中对比学习的性能可能受到限制。

在对比学习中，负样本的质量和数量是制约最后训练效果的关键，如果在训练的批次中出现假负样本，将会影响最终的网络性能。同时，在对比学习研究中经常会遇到崩塌问题，避免神经网络崩塌的最好办法就是同时满足一致性和均匀性。均匀性有助于学习到可分离的特征，但是过度追求均匀性，将会导致一些语义相似的样本特征在一定程度上互相远离。

2.3 生成式融合

近年来，大语言模型（large language model，LLM）（Devlin等，2019；Radford等，2018）取得了显著的研究进展。通过不断扩大数据规模与模型规模，这类大语言模型催生出一系列惊人的能力，典型能力包括：指令遵循能力、上下文学习以及思维链。尽管大语言模型在绝大多数自然语言处理任务上展现出较强的零样本/少样本推理性能，但由于其仅能理解离散的文本信息，因此天生对视觉信息“视而不见”。与此同时，大视觉模型（large vision model，LVM）（Dosovitskiy等，2021；He等，2022）具备优秀的视觉感知能力，但在推理能力方面普遍弱于大语言模型。如何进一步跨越视觉—语言的多模态鸿沟，将图像中丰富的视觉信息转化为符合人类认知逻辑的自然语言描述，是计算机视觉迈向具身智能领域的必要趋势。

正是基于LLM和LVM的能力互补性，催生了多模态大语言模型（multi-large language model，MLLM）（Radford等，2021）这一全新研究领域，MLLM的兴起标志着信息融合技术进入新纪元。其核心思想是：将视觉、音频等非语言模态信息在词元层面实现统一表示，即“翻译”为大语言模型可理解的表示形式。同时，大语言模型作为跨模态理解与生成的枢纽进行自回归建模，并预测下一个词元的分布，从而完成跨模态理解与生成。代表性模型如国际商业版Gemini-3-Pro、GPT-4V等和国内Qwen3-VL、ERNIE-4.5-VL均展示了令人印象深刻的涌现能力：模型不仅能执行训练中见过的任务，还能通过上下文学习完成未显式微调的新任务。Chen等人（2024a）提出了一种双层次的视觉知识增强多模态大型语言模型，通过在细粒度和高级语义层面整合视觉知识提高模型的性能。同时将软提示方法与图像标签高级语义相结合，以减轻不完美预测标签的影响。Dong等人（2024）提出了一种擅长自由格式文本图像合成和理解的视觉语言模型InternLM-XComposer2，该模型能根据多输入创建定制内容，同时采用部分LoRA（low-rank adaptation of large lan-

guage model)方法调整图像标记参数,以保持语言知识完整,实验显示其性能与GPT-4V和Gemini Pro相当或更佳。Wu等人(2024)提出了一种任意多模态大语言融合方法,通过将LLM与多模态适配器和不同解码器连接,使其能感知输入并以任意组合生成文本、图像、视频和音频输出。并且利用现有高性能编码器和解码器,NExT-GPT仅需对少量参数(1%)进行调优,有利于低成本训练和扩展。此外,通过引入模态切换指令,使NExT-GPT具备复杂跨模态语义理解能力。

生成式融合的优势在于同时处理多种模态信息、执行多种任务,避免了单独设计的碎片化融合模块。但同时其计算开销巨大,可解释性较差,并且其精确性、可靠性存疑。三者并非互斥,而是互补关系。注意力机制擅长建模细粒度、动态的跨模态交互,适用于需要精确定位与对齐的任务(如视觉定位)。对比学习在弱监督语义对齐场景中表现卓越,是跨模态检索与预训练的关键技术。生成式融合则以统一的生成框架实现复杂的跨模态推理与创作,代表了通用模型的发展方向。在多模态融合系统中往往三者并用,以对比学习进行大规模预训练获得语义对齐的初始表示,以注意力机制在微调阶段实现任务特定的动态融合,最终以生成式框架进行多样化格式的输入输出。

但是由于生成式大模型本身固有的非确定性、“幻觉”和提示词注入等缺陷,可能导致生成式多模态融合对同一问题生成不同的回答;模型可能产生无意义或者不正确的信息;容易受到攻击者构造的提示词注入风险。

3 典型多模态数据集

COCO数据集(common objects in context)(Lin等,2015)(http://cocodataset.org/#home)。包含Train2017、Val2017和Test2017 3个数据子集,设有汽车、自行车、动物、雨伞和运动器材等80个类别,共330 K图像,其中200 K幅图像具有目标检测、分割和字幕任务的标注。COCO数据集提供了标准化的评估指标,例如用于目标检测的平均精度均值(mean average precision,mAP)以及用于分割任务的平均召回率(mean average recall,mAR),使其适用于比较模型性能。

LAION-400M数据集(large-scale artificial intelligence open network-400M)(Schuhmann等,2021)(https://laion.ai/blog/laion-400-open-dataset)。包含约4亿个图像—文本对,能够实现多组K近邻(K-nearest neighbor,KNN)索引,实现数据集中的快速搜索,并且其img2dataset库能够从URL列表中高效爬取和处理数亿幅图像及其元数据。

Visual Genome数据集(Krishna等,2017)(https://homes.cs.washington.edu/~ranjay/visualgenome/api.html)。美国斯坦福大学视觉实验室的Visual Genome数据集专注于视觉与语言的深度关联,包含超过10万幅图像,每幅图像平均配有1.5个区域描述、对象属性以及对象间关系标注。该数据集具有细粒度的视觉关系标注,为场景图谱构建和视觉问答等相关研究提供丰富素材。

MS COCO Captions数据集(Chen等,2015)(https://cocodataset.org/#home)。MS COCO Captions是建立在微软MS COCO数据集基础上的大型图像数据集。包含超过33万幅日常场景图像,每幅图像提供了至少5段独立的人工文本描述,从而确保了语言的多样性与上下文的准确性。

Conceptual Captions数据集(Sharma等,2018)(https://github.com/google-research-datasets/conceptual-captions)。由谷歌发布的Conceptual Captions集是一个包含约330万图像—文本对的数据集,该数据集的描述文本来自数十亿个互联网网页,通过设置筛选属性自动化清洗和过滤,并保留经过多重筛选的图像—文本描述对,使其成为大视觉语言模型预训练的重要数据资源。

Flickr30k数据集(Plummer等,2015)(https://github.com/bryanplummer/flickr30k_entities)。源于Flickr图像平台的Flickr30k数据集包含3万幅图像,涵盖了多样化的日常生活场景,每幅图像配有5个英文描述句子,以提供丰富的视觉内容和上下文信息,主要用于图像描述生成、视觉问答和图像检索等任务的研究。

此外,针对航空航天等用于航拍场景分类、目标检测与识别等应用的大模型数据集主要包括:武汉大学和华中科技大学联合发布的AID(aerial image dataset)数据集(Xia等,2016)(https://hyper.ai/cn/datasets/5446),包含机场、裸地、港口和工业区等30种航空场景;西北工业大学发布的NWPU VHR-10数

据集(Cheng等,2014)(https://github.com/Gaoshuaikun/NWPU-VHR-10),包含飞机、船只、港口、桥梁和油罐等10种类别、800幅航空目标的遥感图像,常用于目标检测算法评估;由武汉大学等机构发布的DOTA(dataset for object detection in aerial images)数据集(Xia等,2018)(https://captain-whu.github.io/DOTA),包含大量航拍图像和密集的飞机、港口等多类别多角度目标标注,常用于目标检测与识别等技术研究。

由中国科学技术大学与北方电子设备研究所、百度等合作构建的Anti-UAV数据集(Jiang等,2021)(https://github.com/ucas-vg/Anti-UAV),主要用于反无人机检测、识别与跟踪任务。该数据集包含红外与可见光双模态图像序列,覆盖天空、云层、建筑背景以及远距离目标等多种复杂环境。

4 多模态信息融合在军事领域的应用

跨模态信息融合技术在军事领域的应用具有特殊的战略价值。现代战场呈现出传感器多样化、数据海量化以及决策实时化的显著特征,任何单一情报源都难以支撑完整的战场认知。

4.1 战场态势感知与多模态监视

战场态势感知的核心任务是从异构传感器数据中快速识别威胁目标、理解作战意图、预测敌方行动。例如机载传感器雷达—光电—红外融合过程中,雷达提供全天候的目标位置与速度信息,但难以识别目标类型;光电传感器(可见光/红外)提供高分辨率图像,却易受气象条件影响。跨模态融合通过注意力机制将雷达探测到的动目标特征与光电图像中的视觉特征进行关联,可实现目标的精确识别与持续跟踪(Gunning等,2021)。

4.2 多源情报分析与智能处理

现代军事行动依赖于信号情报、图像情报、测量与特征情报等多源情报的协同分析。侦察无人机可同时采集目标区域的通信信号与光电视频。基于对比学习的跨模态模型可将通信内容中的关键词与视频中的视觉目标进行自动关联,实现"听到什么就看到什么"的情报闭环(Chen等,2024b)。同时,文本情报与地理空间数据的融合在作战规划中至关重要。情报报告以自然语言形式存在,而指挥系统需要在地图上标注目标位置。多模态大语言模型能够解析文本中的时空语义,自动转换为结构化地理标注,同时结合历史轨迹数据进行移动趋势预测。

4.3 无人系统协同与自主决策

无人机群的多模态协同感知突破了单平台传感器的物理限制。多架无人机搭载不同传感器(光电、红外、雷达、电子侦察),通过机间数据链共享信息。跨模态融合系统在统一框架下处理来自不同平台、不同模态的异构数据,构建协同态势图(Tao等,2025)。当单架无人机遭遇电子干扰时,融合系统可依赖其他平台的互补信息维持目标跟踪,形成"感知冗余"带来的系统鲁棒性。

5 多模态信息融合的挑战与展望

回顾多模态融合的发展历程,融合的瓶颈正从"表示"转向"交互"。随着对比学习和多模态预训练的成熟,高质量的单模态表示已不再是瓶颈。研究焦点由早期如何将异质模态映射到统一空间,转向如何在表示对齐的基础上设计能够捕捉模态间复杂、动态,甚至矛盾关系的交互机制。

融合架构正从"静态设计"走向"动态自适应"。真实世界中模态的相关性和可靠性随环境动态变化,固定融合策略的局限性日益明显。在军事对抗中,电子干扰可能使特定传感器失效,需根据输入内容与任务上下文动态调整模态激活状态和融合权重,以提升融合系统的鲁棒性。

因果融合模型有望突破当前关联学习的局限。现有方法大多学习模态间的统计相关性,而非因果关系。这导致两个问题:一是模型易学习虚假关联;二是难以泛化到训练分布之外的环境。因此,引入因果推断工具能够使融合模型具备更强的对抗鲁棒性与环境可迁移性。

此外,亟需设计新的多模态融合机制。各模态冗余信息可能造成计算浪费,其相应噪声信息可能影响融合性能。如何利用信息论技术识别并保留互补信息,同时抑制冗余与噪声是多模态融合技术的一个新的发展方向。

6 结 语

本文首先介绍了多模态信息融合的意义和必要性,进而在融合架构层面梳理了从早期融合、晚期融

合到混合融合的演进脉络。同时，详细阐述了基于注意力的交互建模、基于对比学习的语义对齐，以及以大语言模型为载体的生成式融合3种前沿多模态融合方式。进一步介绍了典型多模态数据集及相应应用领域，并且探讨了多模态融合在包括战场态势感知、多源情报分析与无人系统协同等典型军事领域中的应用。此外，基于融合架构、融合模型和计算成本给出了多模态信息融合的3个发展方向。

参考文献（References）

Baltrusaitis T, Ahuja C and Morency L P. 2019. Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2): 423-443 [DOI: 10.1109/TPAMI.2018.2798607]

Brauwers G and Frasincar F. 2023. A general survey on attention mechanisms in deep learning. IEEE Transactions on Knowledge and Data Engineering, 35(4): 3279-3298 [DOI: 10.1109/TKDE.2021.3126456]

Carion N, Massa F and Synnaeve G, 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229

Chen G W, Shen L Y, Shao R, Deng X and Nie L Q. 2024a. LION: empowering multimodal large language model with dual-level visual knowledge//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 26530-26540 [DOI: 10.1109/CVPR52733.2024.02506]

Chen J L, Seng K P, Smith J and Ang L M. 2024b. Situation awareness in AI-based technologies and multimodal systems: architectures, challenges and applications. IEEE Access, 12, 88779-88818 [DOI: 10.1109/ACCESS.2024.3416370]

Chen X L, Fang H, Lin T Y, Vedantam R, Gupta S, Dollar P, et al. 2015. Microsoft COCO captions: data collection and evaluation server [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/1504.00325.pdf

Cheng G, Han J W, Zhou P C and Guo L. 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS Journal of Photogrammetry and Remote Sensing, 98: 119-132 [DOI: 10.1016/j.isprsjprs.2014.10.002]

Chudasama V, Kar P, Gudmalwar A, Shah N, Wasnik P and Onoe N. 2022. M2FNet: multi-modal fusion network for emotion recognition in conversation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 4651-4660 [DOI: 10.1109/CVPRW56347.2022.00511]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA: ACL: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Dong X Y, Zhang P, Zang Y H, Cao Y H, Wang B, Ouyang L K, et al. 2024. InternLM-XComposer2: mastering free-form text-image composition and comprehension in vision-language large model [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/2401.16420.pdf

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: OpenReview.net

Erukude S T, Veluru S R and Marella V C. 2025. Multimodal deep learning: a survey of models, fusion strategies, applications, and research challenges. International Journal of Computer Applications, 187(19): 1-7 [DOI: 10.5120/ijca2025925264]

Gunning D, Vorm E, Wang J Y and Turek M. 2021. DARPA's explainable artificial intelligence program. AI Magazine 40(2): 44-58 [DOI: 10.1609/aimag.v40i2.2850]

Guo R Z, Sun B, Sun X Y, Bu D S and Su S J. 2025. Multimodal fusion object detection method for UAVs under low light conditions. Chinese Journal of Scientifeic Instrument, 46(1): 338-350 (郭润泽, 孙备, 孙晓永, 卜德森, 苏邵璟. 2025. 无人机弱光条件下多模态融合目标检测方法. 仪器仪表学报, 46(1): 338-350) [DOI: 10.19650/j.cnki.cjsi.J2413498]

Hangloo S and Arora B. 2025. Multimodal fusion techniques: review, data representation, information fusion, and application areas. Neurocomputing, 649: #130827 [DOI: 10.1016/j.neucom.2025.130827]

He K M, Chen X L, Xie S N, Li Y H, Dollár P and Girshick R. 2022. Masked autoencoders are scalable vision learners//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15979-15988 [DOI: 10.1109/CVPR52688.2022.01553]

He Y, Xiu J J and Guan X. 2013. Radar data processing with applications. 3rd ed. Beijing: Publishing House of Electronics Industry (何友, 修建娟, 关欣. 2013. 雷达数据处理及应用.3版. 北京: 电子工业出版社)

Jiang N, Wang K R, Peng X K, Yu X H, Wang Q, Xing J L, et al. 2021. Anti-UAV: a large multi-modal benchmark for UAV tracking [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/2101.08466.pdf

Jiao T Z, Guo G P, Feng X Y, Chen Y M and Song J. 2024. A comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications. Computers, Materials and Continua, 80(1): 1-35 [DOI: 10.32604/cmc.2024.053204]

Krishna R, Zhu Y K, Groth O, Johnson J, Hata K, Kravitz J, et al. 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer

Vision, 123(1): 32-73 [DOI: 10.1007/s11263-016-0981-7]

Lan Z, Li Z X, Yan C, Xiang X J, Tang D Q, Zhou H, et al. 2025. Adaptive knowledge distillation with attention-based multi-modal fusion for robust dim object detection. IEEE Transactions on Multimedia, 27: 2083-2096 [DOI: 10.1109/TMM.2024.3521793]

Li J X, Vo D M, Sugimoto A and Nakayama H. 2024. EVCAP: retrieval-augmented image captioning with external visual-name memory for open-world comprehension//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13733-13742 [DOI: 10.1109/CVPR52733.2024.01303]

Li W S, Peng, Y, Zhang M, Ding L, Hu H and Shen L. 2025. Deep model fusion: a survey. IEEE Transactions on Neural Networks and Learning Systems, 37(5): 2008-2024 [DOI: 10.1109/TNNLS.2025.3628666]

Lin T Y, Maire M, Belongie S J, Hays J, Perona P, Ramanan D, et al. 2015. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755

Mushtaq H, Deng X H, Jiang P, Wan S H, Ali M and Ullah I. 2025. GFA-SMT: geometric feature aggregation and self-attention in a multi-head transformer for 3D object detection in autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 26(3): 3557-3573 [DOI: 10.1109/TITS.2024.3520382]

Pan Q, Hu Y M, Lan H, Sun S, Wang Z F and Yang F. 2019. Information fusion progress: joint optimization based on variational Bayesian theory. Acta Automatica Sinica, 45(7): 1027-1223 (潘泉, 胡玉梅, 兰华, 孙帅, 王增福, 杨峰. 2019. 信息融合理论研究进展: 基于变分贝叶斯的联合优化. 自动化学报 45(7): 1027-1223) [DOI: 10.16383/J.AAS.C180029]

Pan Q, Wang Z F, Liang Y and Liu Z G. 2012. Basic methods and progress of information fusion (II). Control Theory and Applications, 29(10): 1233-1244 (潘泉, 王增福, 梁彦, 刘准轧. 2012. 信息融合理论的基本方法与进展(Ⅱ). 控制理论与应用, 29(10): 1233-1244)

Pan Q, Yu X and Cheng Y M. 2003. Essential methods and progress of information fusion theory. Acta Automatica Sinica, 29(4): 599-615 (潘泉, 于昕, 程咏梅. 2003. 信息融合理论的基本方法与进展. 自动化学报, 29(4): 599-615) [DOI: 10.16383/J.AAS.2003.04.016]

Plummer B A, Wang L W, Cervantes C M, Caicedo J C, Hockenmaier J and Lazebnik S. 2015. Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models//Proceedings of 2015 IEEE Conference on IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 2641-2649 [DOI: 10.1109/ICCV.2015.303]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: ICML

Radford A, Narasimhan K, Salimans T and Sutskever I. 2018. Improving language understanding by generative pre-training [EB/OL]. [2026-04-24]. https://cdn. openai. com/research-covers/language-unsupervised/language_understanding_paper.pdf

Rajpurkar P, Zhang J, Lopyrev K and Liang P. 2016. SQuAD: 100, 000+ questions for machine comprehension of text//Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA: ACL: 2383-2392 [DOI: 10.18653/v1/D16-1264]

Ruder S. 2017. An overview of multi-task learning in deep neural networks [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/1706.05098.pdf

Schuhmann C, Vencu R, Beaumont R, Kaczmarczyk R, Mullis C, Katta A, et al. 2021. LAION-400M: open dataset of CLIP-filtered 400 million image-text pairs [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/2111.02114.pdf

Sharma P, Ding N, Goodman S and Soricut R. 2018. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL: 2556-2565 [DOI: 10.18653/v1/P18-1238]

Song K, Kim S H, Tak J, Choi H L and Moon I C. 2016. Data-driven ballistic coefficient learning for future state prediction of high-speed vehicles//Proceedings of the 19th International Conference on Information Fusion. Heidelberg, Germany: IEEE: 17-24

Su Y X, Lan T, Liu Y H, Liu F Y, Yogatama D, Wang Y, et al. 2022. Language models can see: plugging visual controls in text generation [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/2205.02655.pdf

Tang L F, Zhang H, Xu H and Ma J Y. 2023. Deep learning-based image fusion: a survey. Journal of Image and Graphics, 28(1): 3-36 (唐霖峰, 张浩, 徐涵, 马佳义. 2023. 基于深度学习的图像融合方法综述. 中国图象图形学报, 28(1): 3-36) [DOI: 10.11834/jig.220422]

Tao Y, Gao Z, Ye F Q, Xu J B, Song T, Li W D, et al. 2025. Intelligent multimodal multi-sensor fusion-based UAV identification, localization, and countermeasures for safeguarding low-altitude economy [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/2510.22947.pdf

Tzikas D G, Likas A C and Galatsanos N P. 2008. The variational approximation for Bayesian inference. IEEE Signal Processing Magazine, 25(6), 131-146 [DOI: 10.1109/MSP.2008.929620]

Vakil A, Liu J, Zulch P, Blasch E, Ewing R and Li J. 2021. A survey of multimodal sensor fusion for passive RF and EO information integration. IEEE Aerospace and Electronic Systems Magazine, 36(7), 44-61 [DOI: 10.1109/MAES.2020.3006410]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Proceedings of the 31st Inter-

national Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Wang X L, Girshick R and Gupta A. 2018. Non-local neural networks// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: 7794-7803 [DOI: 10.1109/CVPR.2018.00813]

Wu S Q, Fei H, Qu L G, Ji W and Chua T S. 2024. NExT-GPT: any-to-any multimodal LLM//Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR: 53366-53397

Xia G S, Bai X, Ding J, Zhu Z, Belongie S J, Luo J B, et al. 2018. DOTA: a large-scale dataset for object detection in aerial images// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3974-3983 [DOI: 10.1109/CVPR.2018.00418]

Xia G S, Hu J W, Hu F, Shi B G, Bai X, Zhong Y F, et al. 2016. AID: a benchmark dataset for performance evaluation of aerial scene classification [EB/OL]. [2026-04-24]. https://arxiv.org/pdf/1608.05167.pdf

Xue Z H and Marculescu R. 2023. Dynamic multimodal fusion//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2575-2584 [DOI: 10.1109/CVPRW59228.2023.00256]

Zhang Y, Liu Y Q, Li H K, Cheng C and Jia Z Y. 2025. MPFBL: modal pairing-based cross-fusion bootstrap learning for multimodal emotion recognition. Neurocomputing, 658: #131577 [DOI: 10.1016/j.neucom.2025.131577]

作者简介

胡玉梅，女，高级工程师，主要研究方向为多源信息融合和人工智能。E-mail：hym_henu@163.com.

王晓华，男，研究员，主要研究方向为智能信息处理。
E-mail：4283773@qq.com

邓豹，男，研究员，主要研究方向为智能信息处理。
E-mail：hym_henu@126.com

赵洋洋，男，工程师，主要研究方向为故障诊断与智能信息处理。E-mail：1398655303@qq.com

赵艺阳，男，工程师，主要研究方向为人工智能与故障诊断。
E-mail：zhaoyiyang@mail.nwpu.edu.cn