多模态语义通信研究综述

秦志金[1]，赵茭茭[2]，李凡[2]，陶晓明[1]
（1. 清华大学电子工程系，北京 100084；2. 西安交通大学信息与通信工程学院，陕西 西安 710049）

摘 要：随着人工智能与通信的交叉融合，文本、图像、音频、视频等多模态数据处理技术蓬勃发展，模态语义的共享维度被深度挖掘，多模态语义信息的高度抽象、智能简约等特性被充分利用，为语义通信带来了全新的思路和手段。首先，介绍了语义通信的基础理论和分类，分别针对文本、图像、音频、视频综述了单模态语义通信的研究现状；然后，综述了多模态语义通信的研究现状，介绍了多模态数据融合技术和安全语义通信的研究；最后，总结了多模态语义通信面临的挑战。

关键词：语义通信；多模态数据融合；多模态语义通信
中图分类号：TN919.8
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2023105

Survey of research on multimodal semantic communication

QIN Zhijin[1], ZHAO Tantan[2], LI Fan[2], TAO Xiaoming[1]
1. Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
2. School of Information and Communication Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Abstract: With the cross-integration of artificial intelligence and communications, technologies for processing multimodal data such as text, image, audio, and video are booming, the shared dimension of modal semantics is deeply excavated, and the characteristics of multimodal semantic information such as high abstraction, intelligence and simplicity are being fully utilized, which brings new ideas and means to semantic communications. First, the fundamental theories and classifications of semantic communication were introduced, and the research status of single-modal semantic communication was reviewed for text, image, audio, and video respectively. Then, the research status of multimodal semantic communication was reviewed, and multimodal data fusion technology and secure semantic communication were introduced. Finally, the challenges faced by multimodal semantic communication were summarized.

Keywords: semantic communication, multimodal data fusion, multimodal semantic communication

0 引言

过去几十年，通信领域的研究主要集中在如何准确有效地将符号从发送端传输到接收端，也称为语法通信。随着无线通信系统的发展，系统容量逐渐接近香农极限。然而，在万物智能互联的时代，通信的最终目的是交换语义信息而不是准确传输符号。目前，语义通信已经引起了工业界和学术界的广泛关注，有望成为“达意”通信的一种新范式[1]。区别于语法通信，语义通信的主要目的是实现收发端语义信息的准确交互，利用先进的人工智能（AI, artificial intelligence）技术提取出原始数据中与接收端特定的智能任务最相关的信息进行传输，可有效压缩数据冗余，提升信息传输的有效性，减轻网络传输的压力，降低智能任务的处理时延[2]。

随着 6G 技术与 AI 技术的飞速发展和深度融

收稿日期：2023−01−11；**修回日期：**2023−05−06
基金项目：国家自然科学基金资助项目（No.61925105）；清华大学–中国移动联合研究院基金资助项目
Foundation Items: The National Natural Science Foundation of China (No.61925105), Tsinghua University-China Mobile Communications Group Co., Ltd. Joint Institute

合，包括文本、图像、音频、视频等在内的多模态服务必然成为各类场景（如电子医疗、数字孪生、人体感应护理系统、零售店自动结账等）的主流。为了给用户提供比较好的体验，开发面向多模态信号的高效传输和精确处理的系统级通信架构是很有必要的，语义通信有望支持多模态通信架构的实现。高质量的多模态服务可以利用多模态信号的时间、空间和语义关系来保证，在这些跨模态关系中，语义包含反映多模态信号含义的丰富信息，将成为打破模态壁垒的有力媒介，因此，多模态语义通信受到研究者的广泛关注[3]。

本文旨在综述已有单模态语义通信、多模态语义通信的相关工作，介绍多模态数据融合技术，总结现有多模态语义通信面临的挑战，整体框架如图 1 所示。本文的主要贡献如下。

1) 分别针对文本、图像、音频、视频综述单模态语义通信的研究现状。

2) 综述多模态语义通信的研究现状，介绍多模态数据融合技术和安全语义通信研究。

3) 总结多模态语义通信面临的主要挑战，旨在为多模态语义通信后续研究提供可供参考的思路和方向。

本文所述语义通信发展路线如图 2 所示。接下来，对图 2 中每个部分展开详细介绍。

1　语义通信基础理论和分类

1.1　语义通信基础理论

语义的概念起初是在符号学的研究中出现的[4]。Morris[5]把符号学定义为语法、语义和语用的三重组合。语法关注符号（视觉和语言）的形式特征之间的相互关系，而不考虑含义。语义专门研究不同层次的符号含义。语用关注符号系统中符号效用与用户之间的关系。类似于符号的三重定义，Shannon 等[6]确定以下 3 个层面的通信来进一步刻画通信的语法、语义和语用特征[7]。

1) 语法层面：通信符号如何被精确地传输？

2) 语义层面：传输的符号如何准确地传达预期的语义？

3) 有效性层面：接收到的语义如何有效地以预期的方式影响行为？

Carnap 等[8]重新审视了香农研究工作中绕过的语义问题，并对语义信息进行了初步定义。Bao 等[9]首次提出了语义通信的理论以实现语义级别的通信，并定义了语义噪声、语义信道、语义熵和语义信道容量。设信源消息集合为 X，语义信息集合为 W，背景知识为 K，推测为 I，信宿消息集合为 Y。用香农熵 $H(W)$来量化信源的语义信息量，即语义熵。语义熵 $H(W)$和信源熵 $H(X)$之间的关系为

$$H(W)=H(X)+H(W|X)-H(X|W) \tag{1}$$

其中，$H(W|X)$ 衡量编码的语义模糊度，$H(X|W)$ 衡量编码的语义冗余。与经典信息论最大的不同在于，语义信息的衡量基于背景知识和推测决定的逻辑概率，而不是统计概率。

离散无记忆信道的语义信道容量取决于 3 个要素。第一个是 X 和 Y 之间的互信息 $I(X;Y)$，也是经典信息论的信道容量；第二个是用 K_s 和 I_s 进行语义编码时引入的语义模糊度，即 $H_{K_s,I_s}(W|X)$；第三个是接收消息的平均逻辑信息，即 $\overline{H_{K_d,I_d}(Y)}$，由 K_d 和 I_d 决定。如果 $K_s(I_s)$ 和 $K_d(I_d)$ 不匹配，将会产生过多的语义噪声。假设 $K_s=K_d$ 且 $I_s=I_d$，则语义信道容量为

$$C=\sup_{P(W|X)}\left\{I(X;Y)-H(W|X)+\overline{H(Y)}\right\} \tag{2}$$

从式(2)可知，设计合理的语义编解码方案 $P(W|X)$ 对于高效语义通信系统的实现至关重要。语义级别的率失真理论可以为此提供很好的理论指导。

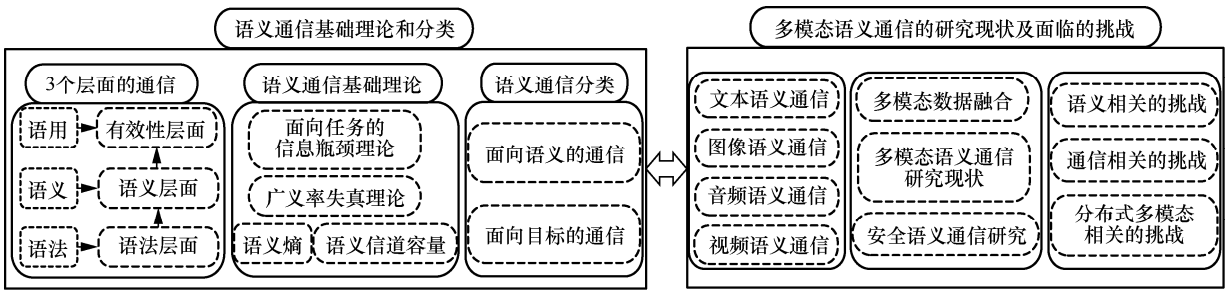

图 1　本文整体框架

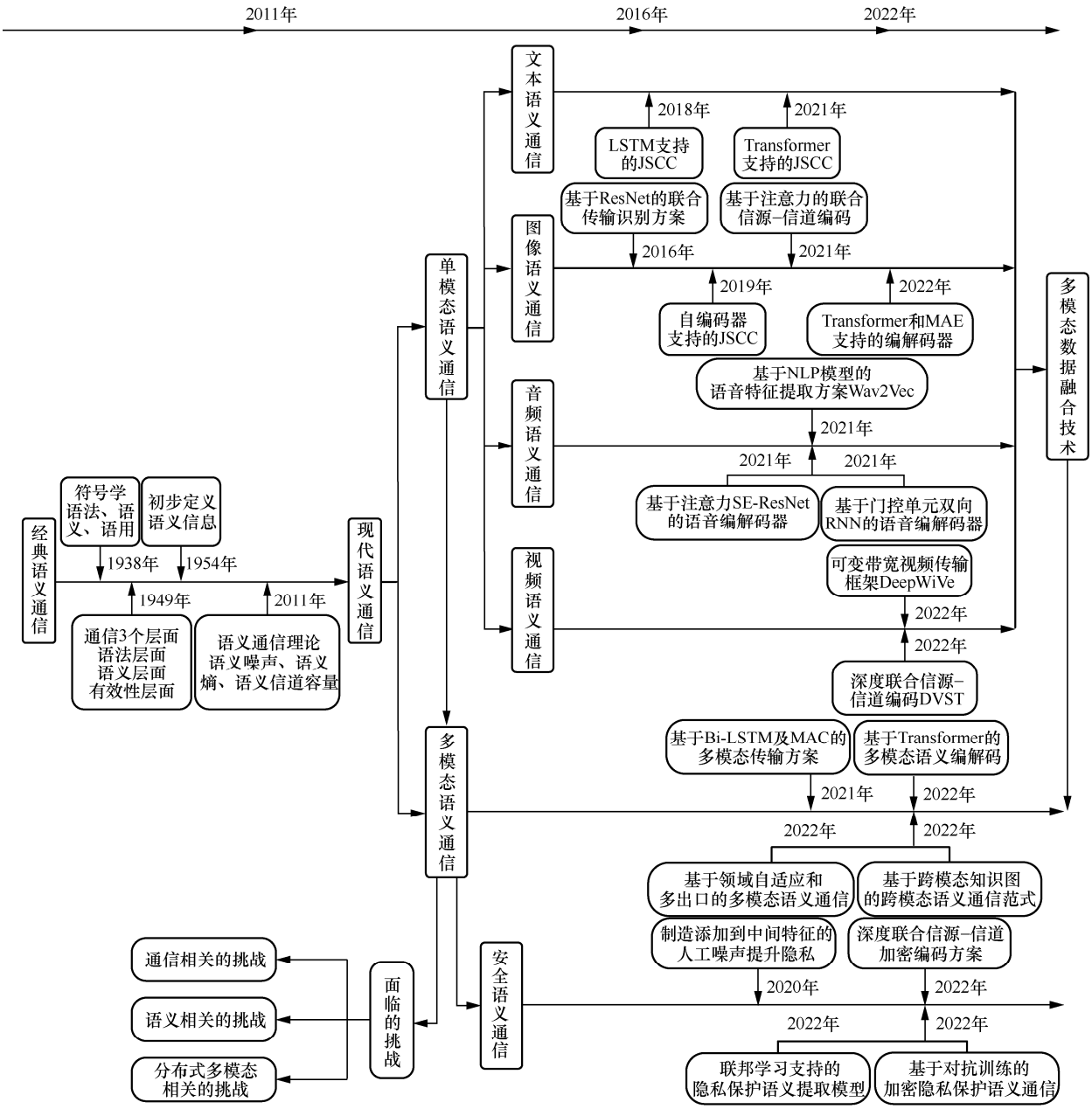

图 2　语义通信发展路线

具体地，广义的率失真理论为

$$\min\left[I(X;Z)+\lambda D(X;Z)\right] \tag{3}$$

其中，$I(X;Z)$ 表示语义特征 Z 保留的关于信源 X 的信息量，衡量语义编码对语义信息的压缩量；$D(X;Z)$ 表示语义特征 Z 和信源 X 的差异，衡量语义编码带来的语义失真量；λ 表示权重因子[10]。

面向任务的信息瓶颈理论可以形式化率失真理论的折中关系[11]，具体表示为

$$\min\left[I(X;Z)-\beta I(Z;Y)\right] \tag{4}$$

其中，Y 为任务标签。特别地，用语义信息的负值 $-I(Z;Y)$ 度量语义失真，失真 $-I(Z;Y)$ 尽可能小意味着语义信息 $I(Z;Y)$ 尽可能大，表示语义特征 Z 中尽可能多地保留任务相关的语义信息[10-11]。

以上关于语义通信的基础理论可以为高效语义通信的设计和实现提供很好的指导，能够根据应用场景和任务需求进行灵活变换，为满足 6G 通信高谱效和高可靠的要求提供新的技术思路[12]。

1.2　语义通信分类

由于强大的 AI 技术，现代语义通信的研究已

经出现在多种应用中。经典通信系统仅关注由 Shannon 等[6]确定的语法层，语义通信则把余下的 2 个更高层融入通信系统的设计中。如图 3 和图 4 所示，语义通信主要分为两类：面向语义的通信（关注语义层面）和面向目标的通信（关注有效性层面）[13]。

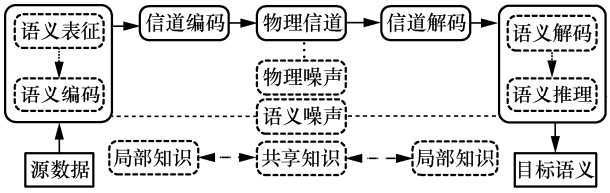

图 3　面向语义的通信

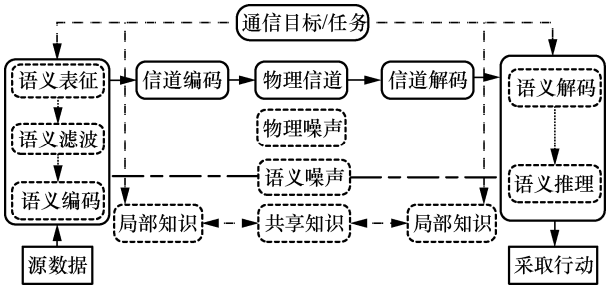

图 4　面向目标的通信

1.2.1　面向语义的通信

不同于忽略传输内容的经典通信系统，面向语义的通信设计中重要的是信源数据语义内容的准确率，而不是与源数据的统计概率相关的平均信息。因此，语义通信系统的主要改变在于发送前和接收后的数据处理阶段。传统的信源编码是寻求一种将信源数据转换为短码的方法，同时，因为发送的消息对潜在的含义视而不见，一个好的信源编码方法意味着它可以处理信源数据更多的可能性。然而，在语义通信中，需要重新定义“信息”，并在编码前引入语义表征模块，负责捕获嵌入在信源数据中的核心信息，过滤不必要的冗余信息，特别地，面向目标的通信中的语义滤波模块负责进一步过滤与下游智能任务无关的信息。很多研究工作把语义表征和语义编码功能集成在一个模块里，称为语义编码，联合发挥与传统通信中信源编码类似的作用。类似地，语义推理和语义解码的联合作用等价于信源解码的作用。

在一般的语义通信场景中，解码是编码的逆过程，可通过 AI 驱动的解码算法实现，如具有强大先验知识的 Transformer 和自编码器（Autoencoder）。语义推理模块基于解码得到的语义信息推理出目标语义或者直接根据语义信息采取行动，完成特定的智能任务。由于语义通信的目标是使接收机成功获取语义信息，因此，可以将联合语义编码和解码过程统一看作“语义提取”。此外，正如人类对话一样，有效的对话要求双方具有关于语言和文化的共同知识。语义通信中，为了确保所有的信源数据能被很好地理解和推理，通信参与方需要及时共享局部知识。如果局部知识不一致，就会产生语义噪声，即使在物理传输没有语法错误的情况下也会导致语义模糊。

1.2.2　面向目标的通信

在面向语义的通信的基础上，面向目标的通信旨在使所涉及的通信参与方能够共同完成一个通信目标或任务。回顾通信的 3 个层面：语法层面、语义层面、有效性层面。在面向语义的通信中，语义提取关注语义信息，而在面向目标的通信中，捕获语用信息是很有必要的。Zhong[4]说明了 3 种信息的相互关系，语用信息可以看作所有能用语法信息传达的语义信息的一部分，且与通信的特定目标相关。对于通信目标频繁改变的各类场景，每次传输时构建局部知识以进一步过滤不相关的语义信息尤其重要，所以，在面向目标的通信中，目标或任务在语义提取中发挥着很重要的作用。面向目标的通信关注有效性层面，在给定有限网络资源的情况下，旨在以预期的方式完成任务，而面向语义的通信关注语义层面的语义信息准确率。此外，类似于面向语义的通信，面向目标的通信中所有通信参与方的局部知识和通信目标需要保持一致，否则，产生的语义噪声会导致任务失败。

1.3　单模态语义通信的研究现状

信源数据主要有文本、音频、图像和视频等各种模态。大多数现有的语义通信研究都围绕上述几种模态展开。其中，可以通过说话或打字来感知的文本是引起最多关注的信源数据类型。在关于文本语义通信的研究中，研究者通常采用语义符号来表征语义，一个语义符号表示单词、短语和句子等数据的子集。如单词“bike”和“bicycle”以及短语“a two-wheeled tool for mobility and transportation”可以映射到同一个语义符号，这也是语义通信可以显著减少带宽的主要原因[14]。然而，这也不可避免地导致一些信息的损失。由此可见，语义编码的压缩比应该由特定应用场景决定。

上述思想可以应用到音频数据传输的语义通信中。近些年，随着声控智能家居应用的兴起，音频通

信不再局限于人与人的对话[15]，语音识别成为一种流行的应用。相比于文本数据，音频数据包含更多的特性，如语速和语气。在一些关于语音识别的语义通信研究中，为了避免其他语音特征的影响，语音信号在进行语义提取之前被转换为文本数据[16]。

此外，通信任务对于视觉数据更加多变，如图像分类、目标识别和视频会议。不同于广义的传统图像和视频压缩及编码，特征提取需要针对通信任务和源数据的特性进行量身定制。以视频会议为例，由于视频会议的背景帧几乎是静止的，Jiang 等[17]把基于关键点的视频恢复技术融入语义通信中，仅关键点（如关于面部表情和行为改变的信息）被实时地编码和传输给接收端，关于背景图片和发言者面部特征的其他信息只是在会议开始时被分享给接收者。这种方式可在保持高水平分辨率的同时实现高压缩比。由于语义通信允许在有限的带宽内传输更多的相关数据，因此，语义通信相比于传统通信可获得更好的性能。

1.3.1 文本语义通信

受到深度学习在自然语言处理（如机器翻译）的启发，Farsad 等[18]率先设计了一个文字传输系统模型，发送者使用有限的比特数通过擦除信道向接收者发送句子。在该模型中，Farsad 等[18]首先使用 Glove[19]把单词表征为嵌入向量，其中，Glove 是可用于提取语义信息的预训练查找表；然后，受到机器翻译中序列到序列学习框架的启发[20-21]，Farsad 等[18]应用基于长短期记忆（LSTM, long short-term memory）的编码器和解码器，把先前估计单词的嵌入向量作为下一步的输入并利用束搜索算法找到最可能的单词序列[22]。这种方式可以在句子恢复过程中嵌入语义信息。然而，诸如 Glove 或 Word2Vec[23] 的词表征模型仅能捕获单词之间的关系，不能描述句法信息[24]。因此，Farsad 等[18]所提模型仅可以描述一个句子中某个单词在另一个单词后出现的概率，很难处理长句子，且未考虑通信环境对文字传输的影响。

面对这个挑战，研究者提出了 Transformer 的新框架并引起了大量的关注，Transformer 可以从整个句子中有效地提取语义信息和语法[24]。具体地，Transformer 网络与允许并行提取句子多个特性的多头注意力机制结合[25]。因此，与诸如 LSTM 这种基于循环神经网络（RNN, recurrent neural network）的结构相比，Transformer 网络具有更低的计算复杂度，能够实现更多的并行计算，同时具备学习输入序列长期依赖关系的能力。

因此，Xie 等[24]提出了基于 Transformer 的联合语义–信道编码方法，用于去除信道噪声及语义噪声对语义通信系统的影响，并将信道模型由擦除信道扩展到加性白高斯噪声（AWGN, additive white Gaussian noise）信道和衰落信道。在低信噪比（SNR, signal-to-noise ratio）范围内，基于 Transformer 的语义通信在 BLEU（bilingual evaluation understudy）和句子相似性度量下具有更明显的优势。然而，Transformer 的注意力结构是固定的。实际上，在一个句子处理系统中，由于多义或噪声干扰，某些单词或短语比其他单词或短语更可能引起语义模糊。考虑到这一点，Zhou 等[26]进一步提出一种灵活的基于通用 Transformer[27]的语义提取方法，这种方法通过在 Transformer 中引入一个自适应循环机制来打破原始的固定结构。

相比于标准 Transformer，通用 Transformer 与自适应计算时间模型[28]结合，可以根据每步预测的停止概率动态调整所需的计算步骤数，处理 RNN 中的每个输入符号。这种动态的每位置停止机制允许基于通用 Transformer 的语义提取可以循环利用自己的机制，实现在不同的周期响应不同的语义信息和变化的物理信道。在仿真中，Zhou 等[26]比较了传统的信源编码和信道编码级联方案、基于通用 Transformer 的语义提取方案和基于标准 Transformer 的语义提取方案的 BLEU 性能。实验结果表明，相比于传统的信源编码和信道编码级联方案，上述 2 种基于 Transformer 的语义通信方案在不同信道条件下可以获得更高的 BLEU 分数。具体地，随着信噪比的变化，2 种方案下 BLEU 的分数趋势是相同的，但是由于自适应循环机制，基于通用 Transformer 的方案得分始终高于基于标准 Transformer 的方案得分。

1.3.2 图像语义通信

对于图像数据，Lee 等[29]考虑了一个简单的图像传输场景，一个物联网（IoT, Internet of things）设备发送图像到服务器完成识别任务，IoT 设备与服务器之间保持直接的点对点无线连接，信道模型为 AWGN 信道和瑞利衰落信道。不同于传统的多个模块级联的通信模型，Lee 等[29]提出了基于深度学习的以识别准确率为性能指标的联合传输–识别方案，采用了性能良好且参数较少的 ResNet 结构[30]。为了在

传输前完成特征提取，ResNet 深度神经网络（DNN, deep neural network）被分割为 2 个部分，前 6 层函数作为发送端的特征提取器，即语义提取器，其余层作为接收端的识别器。

此外，为了完成噪声信道中的自适应语义提取，Lee 等[29]使用 DNN 作为信道编码器和解码器来实现联合语义–信道编码（JSCC, joint semantic-channel coding）。为了证明所提方案的有效性，Lee 等[29]将基于 DNN 的联合传输–识别方案与其他 3 种级联压缩–识别方案分别在模拟和数字传输模式下进行对比。实验结果表明，所提方案在识别准确率和复杂度方面具有最好的性能。通过有效的语义提取，所提方案的识别准确率可以在信噪比高于 0 dB 时达到 0.9。此外，使用训练有素的 DNN 模型，所提方案的运行时间在模拟传输模式下低于 10^{-4} s。然而，所提方案仅能在特定 SNR 下运行。在传统通信系统中，通用信源编码器和解码器能根据 SNR 实现自适应压缩比和信道编码，在给定带宽时达到最优的性能。

为了解决这个问题，Xu 等[31]在有 SNR 反馈时考虑点对点的图像传输系统，将广泛应用于计算机视觉的注意力机制融入特征提取。在 Xu 等[31]的设计中，JSCC 在一个单独的网络中执行，网络包含特征学习模块和注意力特征模块。特征学习模块负责从输入图像中学习特征，然后，注意力特征模块把特征学习模块的输出和 SNR 作为输入，产生一系列可伸缩参数。特别地，特征学习模块和注意力特征模块输出的乘积可以看作特征学习模块输出的滤波版本。解码器也是类似的设计。在仿真中，Xu 等[31]将基于注意力的深度 JSCC 方案与 5 种基本深度 JSCC 方案进行了对比，实验结果显示，Xu 等所提方案的峰值信噪比（PSNR, peak signal-to-noise ratio）是其他基准方案 PSNR 曲线的上包络线，从而证明了基于注意力的深度 JSCC 方案具有更好的鲁棒性、通用性和对宽范围 SNR 的适应性。

此外，考虑到图像数据有更多的空间冗余，Hu 等[32]为图像分类任务提出了资源节约型特征提取模型。在编码过程中，Hu 等[32]使用带有视觉 Transformer 结构[33]的掩码自编码器（MAE, masked autoencoder），并采用一个对称编码–解码器结构。MAE 可以从部分观测中重构一个图像。具体地，首先，一部分原始图像被遮蔽和忽略；然后，在没有被遮蔽的部分嵌入它们在原始图像中的位置信息；最后，送入 Transformer 模块完成图像特征的提取[33]。由于编码器只需要处理未遮蔽块，从而显著减少了内存消耗。相反，解码器的输入是由未遮蔽块的编码特征和遮蔽标记组成的完整标记集，遮蔽标记是一个表明预测块存在的共享学习矢量[32]。

1.3.3　音频语义通信

随着针对文本和图像传输的端到端语义通信系统的发展，Tong 等[15]和 Weng 等[16]进一步研究了面向音频信号传输的语义通信。Tong 等[15]基于深度学习的自然语言处理（NLP, natural language processing）语言模型设计了一个被称为 Wav2Vec 的音频特征提取器。语义编码器由 2 个级联的卷积神经网络（CNN, convolutional neural network）构成，分别被称为特征提取器和特征聚合器[34]。特征提取器负责提取原始音频向量中的粗略音频特征，特征聚合器负责把粗略音频特征融入包含上下文音频特征之间语义关系的高层隐变量[34]。相应地，语义解码器与编码器对称，也是基于 Wav2Vec 结构。然而，在仿真中，语义提取模型在固定信道系数的 AWGN 信道下进行训练，这使在更复杂的信道条件下保证良好的性能变得更具挑战。

类似于文本语义编码器的演进，Weng 等[16,35]进一步将被称为 SE-ResNet 的注意力机制融入特征提取，编码器和解码器由一个或多个顺序连接的 SE-ResNet 模块构成。SE-ResNet 模块中的特征提取是一个具有挤压和激发功能的独立网络单元，负责在训练阶段为与基本信息对应的权重分配较大的值。特别地，挤压操作聚合每个输入特征的二维空间维度，激发操作通过捕获相互依赖关系输出每个特征的注意力因子。同时，采用残差网络缓解由网络深度产生的梯度消失问题。从仿真结果可以看出，Weng 等所提特征提取方法在各种衰落信道和 SNR 下获得了优于传统方法的性能。

Weng 等[36]进一步关注针对英语的语音识别任务，将原始语音样本序列在输入发射机之前转换为频谱。此外，Weng 等[36]引入了单个语音样本序列的转录，每个标记代表字母表中的一个字符或一个单词边界。基于频谱和转录，Weng 等[36]设计了编码器和解码器。语义编码器由 CNN 和基于门控单元的双向 RNN[37]组成。CNN 实现数据压缩，双向 RNN 在传输前提取文本相关的语义特征。信道编码和解码由全连接层实现，语义解码负责将恢复的文

本相关语义特征解码为文本转录。考虑到英文字母表中的字母数量有限，Weng 等[36]设计了一个贪婪的语义解码器，首先，索引所有步骤中的最大概率；然后，使用相应的标记来构建最终的转录。

1.3.4 视频语义通信

除了文本、图像、音频信号，视频逐渐成为人们工作和生活的重要组成部分。Tung 等[38]开发了深度强化学习支持的端到端可变带宽视频传输框架 DeepWiVe，仿真结果表明，DeepWiVe 的多尺度结构相似性指标测度（MS-SSIM, multi-scale structural similarity index measure）性能在所有信道条件下平均优于 H.264 视频压缩和低密度奇偶校验码高达 0.046 2，同时平均优于 H.265+LDPC 高达 0.005 8。Wang 等[39]设计了一类新的深度联合信源–信道编码方案 DVST（deep video semantic transmission），实现视频的无线信道端到端高效传输。整个 DVST 的设计以感知质量和机器视觉任务性能为指标，以最小化端到端的传输率失真为目标。实验结果表明，在标准视频源测试序列和各种通信场景下，DVST 方案的性能优于传统的无线视频编码传输方案，由于它具备视频内容感知和机器视觉任务集成的能力，因此，可以支持未来的语义通信。

Jiang 等[17]提出了一种带有新颖语义错误检测器的视频会议语义传输方案，发言者的照片作为先验信息被共享以帮助构建发言者的面部表情。Jiang 等[17]提出的方案大大降低了对无线传输资源的要求。Tao 等[40]开发了一种移动视频传输框架来保证体验质量，通过建立一个大的数据集来寻求主观体验质量得分和神经网络参数之间的关系以引导语义视频传输。Fried 等[41]提出了通过编辑文本来编辑谈话视频。Tandon 等[42]提出了仅传输文本而非视频的方案，大大降低了网络流量。

2 多模态语义通信的研究现状及面临的挑战

由于多义性和模糊性两大语义问题，针对单模态的语义通信系统很难满足多模态服务的可靠性要求。多义性问题为如何获取真正的语义。文本、图像、音频、视频等模态的源信号本质上是多义的。如果没有相关的背景知识或上下文信息，很难识别源信号试图表达的含义。因此，对于语义编码器而言，仅提取出明确的显示语义是不充分的，应高度关注揭示潜在真实含义的隐含语义。语义模糊性问题为如何精确解释语义。由于语义编码器和无线信道不可避免地存在噪声，恢复信号表达的语义也许无法精确解释发送者真正的语义。同时，由于语义特征的数据量远小于源信号，因此即使很少的比特级传输错误也会导致严重的语义失真。对于语义解码器而言，不仅需要处理比特错误，更重要的是能够精确恢复发送者的预期语义。

由于单模态信号的语义很难克服以上问题，充分利用多模态信号的有效语义，设计针对多模态数据的高效语义编解码器是很有必要的。此外，为了满足用户的极致沉浸式体验，非常有必要开发一种系统级通信架构，以实现文本、图像、音频、视频等多模态数据的精确处理和高效传输。在这种环境下，多模态语义通信应运而生。它可以利用多模态信号的时间、空间和语义联系保证高质量的多模态服务，成为打破模态壁垒的强有力范式，同时，多模态数据融合技术的快速发展将为此范式的成功建立提供强有力的支持。

2.1 多模态数据融合技术

多模态数据是指对于同一个现象，通过不同领域或视角获取到的数据，一般包括文本、图像、音频、视频等。获取这些数据的每一个领域或视角被叫作一个模态[43-44]。由于自然现象的丰富特性，很少有单一的模态提供对感兴趣现象的完整知识。由不同领域或视角获取的关于同一个现象的多模态数据的可用性，为提升任务性能引入了新的自由度[45]。多模态数据融合是指以相关特征或中间决策的形式对不同模态数据的信息进行组合。

不失一般性，图 5 为双模态融合网络的通用结构。多模态数据集由输入输出数据对 $(x, y; z)$ 组成，其中，x 和 y 分别表示 2 种模态，z 表示监督标签；函数 $f(x)$ 和 $g(y)$ 分别以 x 和 y 为输入，输出 $\hat{z}_x$ 和 $\hat{z}_y$，作为真实标签 z 的估计值；函数 f 和 g 分别由 M 层和 N 层组成，子函数表示为 f_l 和 g_l，第 l 层的输出分别为 $x_l = (f_l \circ \cdots \circ f_1)(x)$ 和 $y_l = (g_l \circ \cdots \circ g_1)(y)$。在标准神经网络中，子函数为卷积、池化、乘矩阵和非线性等，这些子函数的输出就是进行跨模态融合的特征。那么，跨模态融合需要解决的问题就是融合哪些特征以及怎样融合这些特征[46]。

建立好的融合结构需要找到多个单模态数据融合的合适位置[47]。按照融合位置的深浅可以将融合方法分为早期融合和后期融合，分别融合低层特征和预测层特征。后期融合在很多情况下的表现优于早期融合[48]。后期融合被定义为多个单模态分支

最终得分的组合。这种组合可以是加权得分平均[49]、双线性乘积[50]或者更加鲁棒的秩最小化[51]。

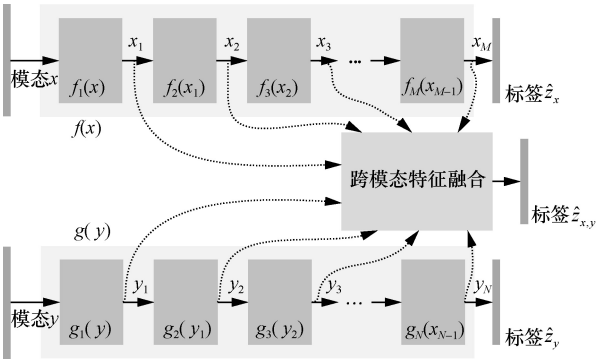

图 5　双模态融合网络的通用结构

多模态融合的另一种架构基于注意力机制[52-54]。注意力机制采用一个额外的神经网络严格地选择某些特征或为原始神经网络中的不同特征分配不同的权重。视觉注意力机制包括多模态双线性池[55]、堆叠注意力网络[56]和自底向上/自顶向下注意力机制[57]。不同于视觉注意力机制，应用于视觉问答（VQA, visual question answering）任务的共同注意力机制同时建模视觉注意力和问题注意力，利用图像和问题的对称性来实现图像表征引导问题注意力和问题表征引导图像注意力[58]。不同于使用浅层模型，由深度级联的模块化共同注意力层组成的深度模块化共同注意力机制在 VQA 任务中表现更佳[59]。

双重注意力网络联合利用视觉和文本注意力机制捕获视觉和语言之间的相互作用，允许视觉和文本注意力在协作推理的过程中相互引导，通过关注图像和句子的共享语义来估计它们之间的相似性，从而有利于 VQA 任务的执行[60]。

2.2　多模态语义通信研究现状

Xie 等[61]以 VQA 为例研究了面向任务的多模态数据传输的语义通信系统。在一个 VQA 任务中，一些用户发送图像，另一些用户发送文本来查询图像信息，在接收端获得回答。Xie 等[61]考虑了带有一个图像发送机、一个文本发送机和一个接收机的简单通信场景。类似于上述针对图像和文本的语义通信研究工作，Xie 等所提图像发送机采用 ResNet-101 网络和预训练的 ImageNet[62]，而所提文本发送机采用双向长短期记忆（Bi-LSTM, bi-directional long short term memory）网络。尽管如此，解码器的设计依然没有得到很好的研究。因为来自 2 个用户的语义信息是相关的，解码器需要融合文本和图像的语义信息同时回答视觉问题。

为了解决这个问题，Xie 等[61]采用记忆力、注意力和合成（MAC, memory, attention, and composition）神经网络[63]来处理相关数据。具体地，每个 MAC 区由控制单元、读取单元和写单元组成。首先，控制单元基于接收到的文本语义信息通过注意力模块生成查询；然后，读取单元接收查询并通过另一个注意力模块从图像语义信息中搜索相应的答案；最后，写单元融合信息并输出问题的预测答案。

此外，Xie 等[64]基于 Transformer 统一图像发送器和文本发送器的语义编码结构。同时，Xie 等[64]提出一个新的语义解码网络，由查询模块和信息融合模块构成。查询模块采用逐层 Transformer，由 Transformer 编码层和 Transformer 解码层构成。不同于经典 Transformer，逐层 Transformer 首先把每个编码层的输出表征作为每个解码层的输入；然后，解码层利用文本信息中更多的关键词和图像信息中的相应区域；最后，融合模块融合这 2 个信息来获取回答。

尽管如此，上述工作依然要求为每个任务训练模型，这限制了它们的应用。因此，Zhang 等[65]设计了一个深度学习支持的统一语义通信系统来服务各类传输任务。为了能用一个模型框架联合服务多个任务，Zhang 等[65]采用领域自适应来降低传输开销。此外，由于每个任务有不同的难度，要求不同数量的层，Zhang 等[65]提出了一种多出口结构，为相对简单的任务提供早出结果。Li 等[3]提出了一种跨模态语义通信范式，通过对音频、视频和触觉信号的跨模态融合和处理来提升语义通信系统的可靠性，包括 3 个模块：跨模态知识图，负责提供基本背景知识和信号块来实现编码和解码；跨模态语义编码器，负责推断潜在的隐式语义以减少编码多义性；跨模态语义解码器，负责保证信源信号和恢复信号在比特级别和语义级别上的一致性，减少解码模糊性。Luo 等[66]通过考虑无线信道传输的性质，提出了一种全新的基于多用户语义通信系统的多模态数据融合方案，即信道级信息融合。在所提方案中，Luo 等[66]将无线信道作为融合多模态数据的媒介且把接收信号看作融合信息，因此，在接收端不需要执行多用户信号检测就能恢复出语义信息。此外，Luo 等[66]设计了语义预编码方案来降低无线信道在融合中的随机效应。在仿真中，Luo 等[66]利用包含 RGB 图像和红外图像 2 种模态的语义分割例子

来证明所提信道级信息融合的可行性和有效性。

多模态语义通信的研究依然处于初级阶段，但是，语义通信在支持各类应用的多模态数据传输、利用多模态数据融合技术充分挖掘多模态数据之间的相关性来降低传输的数据量以及提升语义通信系统的可靠性方面具有巨大的潜力，为充分利用多模态语义通信技术给用户提供具有极致沉浸式体验的多模态服务提供了可行的思路。

2.3 安全语义通信研究

随着通信网络的持续发展，安全已经成为一个重要的课题。作为6G网络的新核心范式，语义通信系统在设计时也需要考虑安全性问题，以满足6G通信网络的强安全要求。无论是单模态还是多模态语义通信，安全问题都不容忽视。联邦学习、对抗学习、添加人工噪声和语义加密等技术有助于构建有效的安全机制来保证语义通信的安全。

相比于传统的通信方法，语义通信能够提升传输中的隐私性和安全性，因为通信参与方仅交换根据通信任务提取的语义信息而不是完整的信源数据，这在很大程度上加强了网络的安全[1]。然而，旨在提升语义提取模型泛化能力的通用知识库的构建引发了隐私问题。因此，Yang等[67]提出了联邦学习支持的语义提取模型训练方案，以隐私保护的方式提升了模型性能。具体地，Yang等[67]首先根据终端设备的接入点和传输要求将它们聚集到不同的组；然后，被分组的终端设备在边缘服务器的调度下，基于各自组的共享背景知识参与特定语义提取任务的预训练或微调。由于不同的组用不共享的知识背景为一个共同的通信目标进行模型参数交换和联邦聚合，因此，在保证模型参数质量的同时保护了隐私。

由于深度神经网络复杂的决策过程易受对抗输入的影响，对抗扰动通过欺骗深度神经网络做出错误任务决策，引发深度神经网络支持的语义通信的安全威胁[68]。数据隐私和语义隐私同等重要，因此，Zheng等[69]首先介绍了衡量数据隐私泄露和语义隐私泄露的2个新指标，具体地，数据隐私泄露用互信息$I(X;Z)$来衡量，其中，X为真实定位，Z为失真定位，语义隐私泄露用真实定位和失真定位的感兴趣点分布概率的互信息$I(P(X),P(Z))$来衡量；然后，提出了语义感知信息论隐私方案来保护数据隐私和语义隐私，同时保留语义感知的数据效用。

随着网络环境的日益复杂，负责复杂模型训练的服务器不总是可信的，这意味着它们对隐私信息是诚实但好奇的。边缘智能协作中，原始数据不会离开边缘设备，只有中间特征被传输和进行进一步的处理。一般而言，通过模型反演和属性推断型攻击能够从这些中间特征中重建出一些隐私数据。所以，如果接收端的解码服务器是诚实但好奇的，就可以由接收到的中间特征通过模型反演和属性推断无差错地重构出原始数据，造成隐私数据泄露。3种可能的隐私保护方法被用来保护隐私免受不可信服务器的侵害[70]，第一种方法是设计能够增加关于隐私信息不确定性（熵）的损失函数，同时减少（或折中）主要任务的错误；第二种方法是制造可以添加到中间特征的噪声来提升隐私[71]；第三种方法是利用对抗学习策略[72]来保护隐私，判别器尝试从中间特征中推测隐私信息，而生成器尝试创建保护它的特征。

此外，加密方法也被用来保护语义通信的安全，Tung等[73]针对无线图像传输，首次提出了可以防止窃听者的深度联合信源–信道编码方案，被称为深度联合信源–信道加密编码。Tung等所提方案不仅保留了深度联合信源–信道编码的有利特性，还提供了针对来自窃听者的选择明文攻击的安全性。具体地，通过在发送端编码模块后进行加密以及接收端解码模块前进行解密的方式来保护语义通信安全。Luo等[74]提出了用于隐私保护的加密语义通信系统，同时考虑了模型的通用性和保密性。通用性体现在所提出的加密语义通信系统的所有网络模块都是基于共享数据库进行训练的，适用于实际场景中的大规模部署，而保密性通过对称加密来实现。基于对抗训练，Luo等[74]设计了对抗加密训练方案以保证加密和非加密2种模式下语义通信的精确度，在训练过程中，合法接收者和非法攻击者同时最小化重构误差，因此，通过交替更新加密器的方式可以减少接收者的重构误差，但这会增加攻击者的重构误差。

3 多模态语义通信面临的挑战

6G网络中，多模态语义通信可以应用在各类以人为中心的数据密集型、计算密集型、时延敏感型、安全敏感型场景中，如元宇宙。元宇宙被设想为未来的互联网。正如人们浏览当今互联网的网页一样，未来人们可以通过头戴式显示器探索元宇宙的虚拟世界，或者通过增强现实眼镜在增强的物理世界中畅游。元宇宙是通过将虚拟世界和物理世界进行同步形成的，其结果是一个人在虚拟域和物理域中的行为有着千丝万缕的联系。由AI技术、边

缘智能技术、虚拟/增强现实技术、区块链技术等驱动，成功实施元宇宙所需的以用户为中心的体验质量、时延、能耗、安全指标需要多模态语义通信网络的快速发展。但是，多模态语义通信的发展依然面临很多挑战，包括语义相关的挑战、通信相关的挑战、分布式多模态相关的挑战等。

3.1　语义相关的挑战

基于深度学习的多模态语义通信的性能优于传统通信，尤其是在低 SNR 范围内。然而，基于深度学习的语义提取架构依然存在一些固有的限制。首先，深度学习范式的损失函数对于收发端的后向传播必须是可微的[75]。因此，所有上述研究依然采用深度学习中常用的损失函数（如交叉熵和最小均方误差）来训练神经网络，这使现有的研究工作与所需的多模态语义通信相去甚远。更重要的是，在端到端架构中，语义和信道编解码需要联合训练，语义提取和恢复被看作一个黑盒[76]，因此，基于深度学习的语义提取缺乏可解释性，它的有效性很难衡量。

作为一个有前途的范式，强化学习可以被用来解决多模态服务、用户定义、任务特定甚至是不可微任务指标的问题。然而，智能体在学习最优策略的过程中，通过与环境的动态交互和奖励函数的设置，在解决语义指标不可微限制、充分发挥多模态数据减少语义多义性和模糊性优势的同时也增加了训练的复杂度。因此，为一个高维度的多模态任务重新训练一个复杂的模型依然非常具有挑战性。

由于多模态任务的特性已经被嵌入损失函数或长期奖励中，上述基于深度学习/强化学习的编码器和解码器对特定的任务目标是无意识的，因此，上述方法仅触及语义层面。由此可见，多模态语义通信场景中基于深度学习/强化学习的语义提取会在特定任务的信息传输中引入不相关的多模态语义信息。面向任务的多模态语义通信可以使系统触及有效性层面。作为一项广泛应用于自动化 AI 系统的技术，知识库以允许推理的形式表征存储数据。在多模态语义通信系统中，利用知识库量化不同目标或任务的语义信息重要性级别，当任务目标改变时，知识库与多模态信号的公共信息一起指导语义提取。尽管如此，如何很好地建立和维护这样一个面向多模态任务的知识库依然需要深入的研究。

不同于单模态语义通信，多模态语义通信中涉及对多模态信号语义特征的提取和融合。虽然现有多模态数据融合技术的巨大成功为多模态语义通信提供了强有力的支持，但是，如何结合多模态语义通信系统的特点打破模态壁垒，在发送端通过多模态语义编码模块的设计实现对多模态信号语义特征的高效提取，在接收端通过多模态语义解码模块的设计实现对多模态语义特征的高效解码和融合，完成特定的多模态任务，依然是一项非常具有挑战性的工作。

3.2　通信相关的挑战

虽然传统通信和多模态语义通信系统使用不同的机制来编码和解码信息，但是，两者面临着相同的通信约束，如不可预测的信道条件、有限的传输和处理资源。然而，不同于传统通信，多模态语义通信需要解决现代通信系统中的新挑战，包括性能分析、资源分配和网络。

传统的通信系统中，信源编码将数据编码为最优长度的符号序列，信道编码给序列添加冗余符号来检测并恢复无线传输中损坏的数据。使用性能分析方法，可以进一步深入理解无线环境和编码机制，从而更好地指导系统设计。在多模态语义通信系统中，信源和信道编码在 AI 技术的支持下联合得更加紧密。联合设计和训练信源–信道编码将更有利于基于深度学习的多模态语义通信系统的数据传输。然而，使用目前不能由数学公式显示表达和精确解释的深度学习方法限制了多模态语义通信系统性能分析的实现。多模态语义通信系统设计者必须考虑如何在时变的无线环境与复杂的多模态语义通信机制之间建立联系，从而进一步指导系统的高效设计。

数据传输需要一些资源，如带宽和功率。一方面，传统通信系统中的资源分配框架旨在最小化误比特率、误包率和中断概率等指标。另一方面，多模态语义通信重视比特流背后多模态语义信息的重要性，从而激发了为新的多模态语义通信系统开发新的资源分配框架。一般来说，通过设计资源分配方法建立一个有效的通信系统应该考虑服务质量（QoS, quality of service）和体验质量（QoE, quality of experience）。具体地，QoS 旨在优化传输速率、时延和吞吐量，而 QoE 关注用户满意度、清晰度和流畅度。因此，多模态语义通信系统中的资源分配策略应该考虑多模态语义信息的非均匀分布，把更多的带宽资源分配给具有更多语义信息的多模态数据/智能体，同时分配更多的功率资源来传输包含更丰富语义信息的多模态数据以保证功率的高效使用。

对于多模态语义通信网络，无线通信层将会对系统性能产生比端到端通信更大的影响。因为很多

产生多模态数据的异构设备工作在语义通信网络中，在设备硬件和无线环境方面的不同将会给多模态语义通信系统的构建带来挑战，主要体现在不同的设备能力、智能化连接的 IoT 网络、多模态编码和解码方案。此外，AI 技术若应用于语义通信中，其算法复杂度也是必须考虑的核心要素。一方面，由于深度的网络结构和大量的训练数据，AI 技术的算法复杂度较高，需要消耗大量的计算资源。另一方面，由于大量的数据分布在位于网络边缘的终端设备，要完成模型的训练，终端设备和负责 AI 模型训练的云服务器或者边缘服务器需要进行频繁的高代价通信，因此，要保证各类智能任务的实时、高效完成，必须对基于 AI 的信息处理资源和通信资源统筹编排，基于云边端协同的机制，充分利用位于网络边缘的设备的丰富训练数据资源和有限信息处理计算资源，通过单模态和多模态语义通信处理技术，对原始训练数据进行适当的预分析处理，并设计与之匹配的通信资源编排策略，提取和传输最有利于下游智能任务高效准确执行的语义信息。随着网络大规模接入的异构设备数量的日益增加，通过对网络有限的计算和通信资源的统筹编排，实现基于人工智能的信息处理速度与通信速度相匹配，完成多样化新兴应用安全、实时、可靠、高效的性能要求是一个具有挑战性的课题。

3.3 分布式多模态相关的挑战

多模态语义通信通过传输和处理来自多个分布式异构终端的单模态信源信号，产生面向不同应用程序的更具信息性和益处的融合信息。例如，通过融合点云、图像、雷达、声音等，非视距环境中车辆检测和追踪的精确度显著提升。物联网中，多模态传感数据（如温度/湿度传感器、气体传感器、光传感器等）的融合有利于环境的有效监控，如大气环境监测、水质监测等。

但是，无线通信环境中，传感设备的位置通常是分布式的，因此，数据必须传输给一个接收者进行信息融合。信号同步成为一个重要的问题，即来自不同发送者的所有信号必须同时到达接收者，这在很大程度上取决于信号传输时间和无线信道条件。首先，语义源必须协作地传输多模态数据，这意味着一个发送者的信号传输时间必须被其他发送者已知。此外，发送者分布式的地理位置导致传输时延不同，多径传播使这个问题更加复杂。如果不需要的信息在无线信道上融合，来自不同步发送者的信号会干扰其他同步发送者，这种同频干扰不容忽视。此外，这会引起接收者最终结果的模糊性，并降低语义通信的性能。然而，保证同步极其困难，需要精确估计无线信道和信号传输时延。

信息融合中，数据是多模态的且由多个异构设备产生，如摄像机、激光雷达、温度/湿度传感器、智能手机、车辆、无人机、声学传感器、地震传感器等。由于不同传感器产生的数据格式是多样化的，因此，需要为每个单模态传感器设计一个编码器结构，然后，将所有数据转换成相同的格式。此外，语义通信的信息融合需要一个统一的转换标准。每种模态的编码器可以不同，这导致编码器压缩后附加模态数据的维度不一致。尤其在信道级信息融合方案中，由于不需要在接收端检测多用户信号，因此最好使来自不同发送者的发送符号维度相同，这使多模态语义通信变得非常具有挑战性。

总之，多模态语义通信需要 6G 网络能够对多种模态信号进行高效处理和传输，同时，保证产生多种模态数据的多个异构设备工作在一个语义通信网络，如何从设备层面、网络层面、业务层面对网络中的计算和通信资源进行统筹调度，通过多模态语义通信，完成 6G 新兴应用强安全、超低时延、超高可靠、超低能耗、超高体验质量的要求，是一个具有挑战性的课题。现有多模态语义通信的研究还处于初级和快速发展阶段，很多基本概念需要发展和改进。首先，多模态语义基本要素的准确性决定了实际应用中的可靠性，是多模态语义通信中最根本的课题。其次，如何设计有效的多模态语义容错和纠错机制仍是未知数。此外，仍然缺乏可以在资源有限的设备中实现快速的多模态语义信息检测和处理的简单而通用的方案。不同实体之间的语义通信模型很难共享，这加剧了在通信系统中采用多模态语义通信的挑战。同时，语义通信也对网络安全提出了完全不同的要求，这需要建立完善的审查机制来防止语义知识图谱被恶意篡改，以及安全可靠的存储和召回机制来防止用户隐私信息泄露。

随着 5G 和 6G 技术的发展，出现了很多先进的通信技术，如非正交多址[77]、大规模多输入多输出、可配置智能反射面、移动边缘计算[78-79]、联邦学习[80]、可见光通信、毫米波[81-82]等；同时，随着 AI 技术的发展，出现了很多功能强大的神经网络架构，如 CNN、RNN、LSTM 等。这些技术能为应对多模态语义通信中语义相关、通信相关、分布式多

模态相关的挑战提供很大的潜力。因此，如何在将来的工作中把先进的通信技术、功能强大的 AI 技术和多模态语义通信结合起来，为构建高效的多模态语义通信范式提供可行思路非常值得研究。

4　结束语

目前，面向 6G 的多模态语义通信的研究仍处于初级阶段，本文主要围绕多模态语义通信的基础理论、研究现状、关键技术及所面临的挑战 4 个方面展开。首先，介绍了语义通信的基础理论，包括语义熵、语义信道容量、广义的率失真理论、面向任务的信息瓶颈理论，同时，将语义通信分为面向语义和面向目标两类；然后，综述了单模态语义通信的研究现状，分别回顾了文本、图像、音频、视频 4 种常见模态的现有研究工作；其次，综述了多模态语义通信的研究现状，介绍了多模态数据融合技术和安全语义通信研究；最后，总结了多模态语义通信面临的主要挑战，包括语义相关、通信相关和分布式多模态相关的挑战，为多模态语义通信后续可能的研究方向提供可供参考的思路。

参考文献：

[1] QIN Z, TAO X, LU J, et al. Semantic communications: principles and challenges[J]. arXiv Preprint, arXiv: 2201.01389, 2022.

[2] 刘传宏, 郭彩丽, 杨洋, 等. 人工智能物联网中面向智能任务的语义通信方法[J]. 通信学报, 2021, 42(11): 97-108.
LIU C H, GUO C L, YANG Y, et al. Intelligent task-oriented semantic communication method in artificial intelligence of things[J]. Journal on Communications, 2021, 42(11): 97-108.

[3] LI A, WEI X, WU D, et al. Cross-modal semantic communications[J]. IEEE Wireless Communications, 2022, 29(6): 144-151.

[4] ZHONG Y X. A theory of semantic information[J]. China Communications, 2017, 14(1): 1-17.

[5] MORRIS C W. Foundations of the theory of signs[M]. Chicago: University of Chicago Press, 1938.

[6] SHANNON C E, WEAVER W. The mathematical theory of communication[M]. Urbana: University of Illinois Press, 1998.

[7] ZHANG P, XU W, GAO H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks[J]. Engineering, 2022, 8: 60-73.

[8] CARNAP R, BAR-HILLEL Y. An outline of a theory of semantic information[J]. The Journal of Symbolic Logic, 1954, 19(3): 230-232.

[9] BAO J, BASU P, DEAN M K, et al. Towards a theory of semantic communication[C]//Proceedings of 2011 IEEE Network Science Workshop. Piscataway: IEEE Press, 2011: 110-117.

[10] 刘传宏, 郭彩丽, 杨洋, 等. 面向智能任务的语义通信: 理论、技术和挑战[J]. 通信学报, 2022, 43(6): 41-57.
LIU C H, GUO C L, YANG Y, et al. Intelligent task-oriented semantic communications: theory, technology and challenges[J]. Journal on Communications, 2022, 43(6): 41-57.

[11] SHAO J W, MAO Y Y, ZHANG J. Learning task-oriented communication for edge inference: an information bottleneck approach[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(1): 197-211.

[12] 张海君, 陈安琪, 李亚博, 等. 6G 移动网络关键技术[J]. 通信学报, 2022, 43(7): 189-202.
ZHANG H J, CHEN A Q, LI Y B, et al. Key technologies of 6G mobile network[J]. Journal on Communications, 2022, 43(7): 189-202.

[13] CALVANESE S E, BARBAROSSA S. 6G networks: beyond Shannon towards semantic and goal-oriented communications[J]. Computer Networks, 2021, 190: 107930.

[14] SHI G, GAO D, SONG X, et al. A new communication paradigm: from bit accuracy to semantic fidelity[J]. arXiv Preprint, arXiv: 2101.12649, 2021.

[15] TONG H N, YANG Z H, WANG S H, et al. Federated learning for audio semantic communication[J]. Frontiers in Communications and Networks, 2021, 2: 734402.

[16] WENG Z Z, QIN Z J, LI G Y. Semantic communications for speech signals[C]//Proceedings of 2021 IEEE International Conference on Communications. Piscataway: IEEE Press, 2021: 1-6.

[17] JIANG P, WEN C K, JIN S, et al. Wireless semantic communications for video conferencing[J]. arXiv Preprint, arXiv: 2204.07790, 2022.

[18] FARSAD N, RAO M, GOLDSMITH A. Deep learning for joint source-channel coding of text[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 2326-2330.

[19] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1532-1543.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv Preprint, arXiv: 1409.0473, 2014.

[21] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation[J]. arXiv Preprint, arXiv: 1609.08144, 2016.

[22] GRAVES A. Sequence transduction with recurrent neural networks[J]. arXiv Preprint, arXiv: 1211.3711, 2012.

[23] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint, arXiv: 1301.3781, 2013.

[24] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems[J]. IEEE Transactions on Signal Processing, 2021, 69: 2663-2675.

[25] SANA M, STRINATI E C. Learning semantics: an opportunity for effective 6G communications[C]//Proceedings of 2022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC). Piscataway: IEEE Press, 2022: 631-636.

[26] ZHOU Q Y, LI R P, ZHAO Z F, et al. Semantic communication with adaptive universal transformer[J]. IEEE Wireless Communications Letters, 2022, 11(3): 453-457.

[27] DEHGHANI M, GOUWS S, VINYALS O, et al. Universal transformers[J]. arXiv Preprint, arXiv: 1807.03819, 2018.

[28] GRAVES A. Adaptive computation time for recurrent neural networks[J]. arXiv Preprint, arXiv: 1603.08983, 2016.

[29] LEE C H, LIN J W, CHEN P H, et al. Deep learning-constructed joint transmission-recognition for Internet of things[J]. IEEE Access, 2019, 7: 76547-76561.

[30] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 770-778.

[31] XU J L, AI B, CHEN W, et al. Wireless image transmission using deep source channel coding with attention modules[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(4): 2315-2328.

[32] HU Q, ZHANG G, QIN Z, et al. Robust semantic communications against semantic noise[J]. arXiv Preprint, arXiv: 2202.03338, 2022.

[33] HE K M, CHEN X L, XIE S N, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15979-15988.

[34] SCHNEIDER S, BAEVSKI A, COLLOBERT R, et al. Wav2Vec: unsupervised pre-training for speech recognition[J]. arXiv Preprint, arXiv: 1904.05862, 2019.

[35] WENG Z Z, QIN Z J. Semantic communication systems for speech transmission[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2434-2444.

[36] WENG Z Z, QIN Z J, LI G Y, Semantic communications for speech recognition[J]. arXiv Preprint, arXiv: 2107.11190, 2021.

[37] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[38] TUNG T Y, GÜNDÜZ D. DeepWiVe: deep-learning-aided wireless video transmission[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(9): 2570-2583.

[39] WANG S, DAI J, LIANG Z, et al. Wireless deep video semantic transmission[J]. arXiv Preprint, arXiv: 2205.13129, 2022.

[40] TAO X M, DUAN Y P, XU M, et al. Learning QoE of mobile video transmission with deep neural network: a data-driven approach[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(6): 1337-1348.

[41] FRIED O, TEWARI A, ZOLLHÖFER M, et al. Text-based editing of talking-head video[J]. ACM Transactions on Graphics, 2019, 38(4): 1-14.

[42] TANDON P, CHANDAK S, PATARANUTAPORN P, et al. Txt2Vid: ultra-low bitrate compression of talking-head videos via text[J]. arXiv Preprint, arXiv: 2106.14014, 2021.

[43] 赵亮. 多模态数据融合算法研究[D]. 大连: 大连理工大学, 2018.
ZHAO L. Research on multimodal data fusion algorithm[D]. Dalian: Dalian University of Technology, 2018.

[44] 任泽裕, 王振超, 柯尊旺, 等. 多模态数据融合综述[J]. 计算机工程与应用, 2021, 57(18): 49-64.
REN Z Y, WANG Z C, KE Z W, et al. Survey of multimodal data fusion[J]. Computer Engineering and Applications, 2021, 57(18): 49-64.

[45] LAHAT D, ADALI T, JUTTEN C. Multimodal data fusion: an overview of methods, challenges, and prospects[J]. Proceedings of the IEEE, 2015, 103(9): 1449-1477.

[46] PEREZ-RUA J M, VIELZEUF V, PATEUX S, et al. MFAS: multimodal fusion architecture search[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 6959-6968.

[47] VIELZEUF V, LECHERVY A, PATEUX S, et al. CentralNet: a multilayer approach for multimodal fusion[J]. arXiv Preprint, arXiv: 1808.07275, 2018.

[48] SNOEK C G M, WORRING M, SMEULDERS A W M. Early versus late fusion in semantic video analysis[C]//Proceedings of the 13th Annual ACM International Conference on Multimedia. New York: ACM Press, 2005: 399-402.

[49] NATARAJAN P, WU S, VITALADEVUNI S, et al. Multimodal feature fusion for robust event detection in web videos[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 1298-1305.

[50] BEN-YOUNES H, CADENE R, CORD M, et al. MUTAN: multimodal tucker fusion for visual question answering[C]//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 2631-2639.

[51] YE G N, LIU D, JHUO I H, et al. Robust late fusion with rank minimization[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3021-3028.

[52] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[J]. arXiv Preprint, arXiv: 1406.6247, 2014.

[53] WANG F, JIANG M Q, QIAN C, et al. Residual attention network for image classification[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 6450-6458.

[54] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.

[55] KIM J H, ON K W, LIM W, et al. Hadamard product for low-rank bilinear pooling[J]. arXiv Preprint, arXiv: 1610.04325, 2016.

[56] YANG Z C, HE X D, GAO J F, et al. Stacked attention networks for image question answering[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 21-29.

[57] ANDERSON P, HE X D, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6077-6086.

[58] LU J S, YANG J W, BATRA D, et al. Hierarchical question-image co-attention for visual question answering[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 289-297.

[59] YU Z, YU J, CUI Y H, et al. Deep modular Co-attention networks for visual question answering[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 6274-6283.

[60] NAM H, HA J W, KIM J. Dual attention networks for multimodal reasoning and matching[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2156-2164.

[61] XIE H, QIN Z, LI G Y. Task-oriented semantic communications for multimodal data[J]. arXiv Preprint, arXiv: 2108.07357, 2021.

[62] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[63] HUDSON D A, MANNING C D. Compositional attention networks for machine reasoning[J]. arXiv Preprint, arXiv: 1803.03067, 2018.
[64] XIE H Q, QIN Z J, TAO X M, et al. Task-oriented multi-user semantic communications[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(9): 2584-2597.
[65] ZHANG G, HU Q, QIN Z, et al. A unified multi-task semantic communication system with domain adaptation[J]. arXiv Preprint, arXiv: 2206.00254, 2022.
[66] LUO X W, GAO R B, CHEN H H, et al. Multi-modal and multi-user semantic communications for channel-level information fusion[J]. IEEE Wireless Communications, 2022: doi.org/10.1109/MWC.011.2200288.
[67] YANG W, LIEW Z Q, LIM W Y B, et al. Semantic communication meets edge intelligence[J]. arXiv Preprint, arXiv: 2202.06471, 2022.
[68] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers[J]. IEEE Transactions on Wireless Communications, 2022, 21(6): 3868-3880.
[69] ZHENG Z R, LI Z T, JIANG H B, et al. Semantic-aware privacy-preserving online location trajectory data sharing[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2256-2271.
[70] BAJIĆ I V, LIN W S, TIAN Y H. Collaborative intelligence: challenges and opportunities[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 8493-8497.
[71] MIRESHGHALLAH F, TARAM M, RAMRAKHYANI P, et al. Shredder: learning noise distributions to protect inference privacy[C]//Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. New York: ACM Press, 2020: 3-18.
[72] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
[73] TUNG T Y, GUNDUZ D. Deep joint source-channel and encryption coding: secure semantic communications[J]. arXiv Preprint, arXiv: 2208.09245, 2022.
[74] LUO X, CHEN Z, TAO M, et al. Encrypted semantic communication using adversarial training for privacy preserving[J]. arXiv Preprint, arXiv: 2209.09008, 2022.
[75] LU K, ZHOU Q Y, LI R P, et al. Rethinking modern communication from semantic coding to semantic communication[J]. IEEE Wireless Communications, 2023, 30(1): 158-164.
[76] SEO H, PARK J, BENNIS M, et al. Semantics-native communication with contextual reasoning[J]. arXiv Preprint, arXiv: 2108.05681, 2021.
[77] ZHAO T T, LI G B, ZHANG G M, et al. Security-enhanced user pairing for MISO-NOMA downlink transmission[C]//Proceedings of 2018 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2019: 1-6.
[78] ZHAO T T, HE L J, HUANG X Y, et al. QoE-driven secure video transmission in cloud-edge collaborative networks[J]. IEEE Transactions on Vehicular Technology, 2022, 71(1): 681-696.
[79] ZHAO T T, HE L J, HUANG X Y, et al. DRL-based secure video offloading in MEC-enabled IoT networks[J]. IEEE Internet of Things Journal, 2022, 9(19): 18710-18724.
[80] ZHAO T T, LI F, HE L J. DRL-based joint resource allocation and device orchestration for hierarchical federated learning in NOMA-enabled industrial IoT[J]. IEEE Transactions on Industrial Informatics, 2022: doi.org/10.1109/TII.2022.3170900.
[81] LIU Y Q, XU K D, LI J X, et al. Millimeter-wave E-plane waveguide bandpass filters based on spoof surface plasmon polaritons[J]. IEEE Transactions on Microwave Theory and Techniques, 2022, 70(10): 4399-4409.
[82] LIU Y Q, XU K D. Design of millimeter-wave bandpass filter using edge-coupling dual-mode resonator[C]//Proceedings of 2021 IEEE Asia-Pacific Microwave Conference (APMC). Piscataway: IEEE Press, 2022: 154-156.

[作者简介]

秦志金（1989– ），女，山西太原人，博士，清华大学副教授、博士生导师，主要研究方向为语义通信等。

赵茭茭（1991– ），女，甘肃陇南人，西安交通大学博士生，主要研究方向为无线安全传输、移动边缘计算、深度强化学习、联邦学习等。

李凡（1981– ），男，陕西宝鸡人，博士，西安交通大学教授、博士生导师，主要研究方向为基于深度学习的图像视频编码、基于机器学习的图像视频质量评价、图像视频的深度理解和处理等。

陶晓明（1981– ），女，河北石家庄人，博士，清华大学教授、博士生导师，主要研究方向为无线多媒体通信理论及关键技术应用等。