面向 6G 的深度图像语义通信模型

江沸菠[1]，彭于波[1]，董莉[2,3]

（1. 湖南师范大学信息科学与工程学院，湖南 长沙 410081；2. 湖南工商大学长沙人工智能社会实验室，湖南 长沙 410205；
3. 湘江实验室，湖南 长沙 410205）

摘 要：目前的语义通信模型在处理图像数据方面仍有可改善的部分，包括有效的图像语义编解码、高效的语义模型训练和精准的图像语义评估。为此，提出了一种深度图像语义通信（DeepISC）模型。首先采用基于 vision transformer 的自编码器（ViTA）网络实现高质量的图像语义编解码；接着采用自编码器实现信道编解码，保证语义在信道上的传输；然后利用判别器网络（DSN）和 ViTA 的双网络架构协同训练，提高重建图像的语义精度；最后针对不同的下游视觉任务提出不同的图像语义评估指标。仿真结果表明，相较于其他方案，DeepISC 可以更有效地还原传输图像的语义特征，使重建图像在各个下游任务中都展现出与原图像相同或相近的语义结果。

关键词：人工智能；6G；语义通信；图像识别；特征提取

中图分类号：TN929.5

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023050

Deep image semantic communication model for 6G

JIANG Feibo[1], PENG Yubo[1], DONG Li[2,3]

1. School of Information Science and Engineering, Hunan Normal University, Changsha 410081, China
2. Changsha Social Laboratory of Artificial Intelligence, Hunan University of Technology and Business, Changsha 410205, China
3. Xiangjiang Laboratory, Changsha 410205, China

Abstract: Current semantic communication models still have some parts that can be improved in processing image data, including effective image semantic codec, efficient semantic model training, and accurate image semantic evaluation. Hence, a deep image semantic communication (DeepISC) model was proposed. The vision transformer-based autoencoder (ViTA) network was used to achieve high-quality image semantic encoding and decoding. Then, an autoencoder realized channel codec to ensure the transmission of semantics on the channel. Furthermore, the discriminator network (DSN) and ViTA's dual network architecture were used to jointly train, thus improving the semantic accuracy of the reconstructed image. Finally, for different downstream vision tasks, different evaluation indicators of image semantics were presented. Simulation results show that compared with other schemes, DeepISC can more effectively restore the semantic features of the transmitted image, so that the reconstructed image can show the same or similar semantic results as the original image in various downstream tasks.

Keywords: artificial intelligence, 6G, semantic communication, image recognition, feature extraction

收稿日期：2022−11−22；**修回日期：**2023−02−09

通信作者：彭于波，pengyubo@hunnu.edu.cn

基金项目：国家自然科学基金资助项目（No.41904127, No.41604117）；湘江实验室开放基金资助项目（No.6108408DL001, No.6109408DL001）；湖南省教育厅科学研究优秀青年基金资助项目（No.7103408DL001）；湖南省教育厅资助科研项目（No.21A0372）

Foundation Items: The National Natural Science Foundation of China (No.41904127, No.41604117), Open Project of Xiangjiang Laboratory (No.6108408DL001, No.6109408DL001), Project of Outstanding Youth in Scientific Research of Hunan Provincial Department of Education (No.7103408DL001), Scientific Research Fund of Hunan Provincial Education Department (No.21A0372)

0　引言

移动通信系统经过 1G 到 5G 的发展，数据的传输速率实现了巨大的飞跃，但系统容量也逐渐接近香农极限[1]。随着人工智能（AI, artificial intelligence）技术和物联网（IoT, Internet of things）技术的相互融合，各种新的智能应用层出不穷，如自主运输、消费机器人、环境监测和远程医疗等[2]，这些都大大加快了智能物联网（AIoT, artificial intelligence of things）的发展，使万物智联成为时代所趋。为了给人们提供越来越多的智能服务，如智能家居、智能制造和智慧城市等，大量的智能物联网设备被广泛部署。这些设备需要在有限的频谱资源上支持大规模的连接并要求较低的时延，这将导致频谱资源的巨大消耗以及通信能耗的急剧增长。

作为一种新的智能通信范式，6G 语义通信引起了研究者的广泛关注。与传统通信不同，语义通信旨在传输与语义相关的信息，它通过提取数据的语义特征，在保留主要含义的同时进一步压缩数据。因此，语义通信可以有效地减少频谱资源的浪费，并且在恶劣的信道环境中，尤其是在低信噪比（SNR, signal to noise ratio）的环境中仍具有较强的稳健性[1]。

随着深度学习（DL, deep learning）和智能硬件的蓬勃发展，许多研究者已经开始研究基于深度学习的语义通信方法。在面向文本数据的语义通信方面，Xie 等[1]提出一种基于深度学习的语义通信系统用于文本传输。该系统通过恢复句子的含义来最大化系统容量并最小化语义错误，并且使用迁移学习加速模型的训练过程，保证该系统可以适用于不同的通信环境。在文献[1]的基础上，Xie 等[3]提出一种基于深度学习的精简分布式语义通信系统，以实现低复杂度的文本传输，使从 IoT 设备到云/边缘的数据传输可以在语义级别进行，以提高传输效率。在面向语音数据的语义通信方面，Tong 等[4]研究了无线网络中基于音频的语义通信问题，提出一种由卷积神经网络（CNN, convolutional neural network）组成的基于波向量架构的自编码器，该编码器能够以少量数据实现高精度的音频传输。Kotsakis 等[5]研究了在广播场景下，如何使用各种音频模式分类器对广播音频进行语义分析，并提出一种基于有监督训练的广播节目自适应分类策略。

6G 中的一些新型应用，如混合现实和自动驾驶等，将广泛地使用图像数据，因此面向图像的语义通信也引起了研究者的重点关注。Huang 等[6]提出一种基于生成对抗网络的图像语义编码方式，为多媒体语义通信系统设计了一个从粗到细的图像语义编码模型。Patwa[7]等提出了一种压缩视觉数据的方法，以最大限度地提高对目标任务的分析性能。为了在图像无线传输过程中为目标任务保留像素背后的语义信息，Sun 等[8]提出了一种基于像素语义的联合源通道编码方法。Wang 等[9]引入了对抗性损失来优化基于深度学习的联合源通道编码，使其倾向于更好地保留全局语义信息和局部纹理。Wang 等[10]研究了语义先验信息在图像压缩中的重要性，并设计了一个语义先验注意力模块来自适应地增强语义特征。Hu 等[11]提出了一个端到端的语义通信系统框架，显著提高了语义通信系统对语义噪声的稳健性并降低了传输开销。

虽然已有的图像语义通信模型可以保证较好的语义通信结果，但仍有以下 3 个部分可以进一步改善。

1) 图像语义编解码器。基于卷积的图像语义编码器虽然有出色的图像局部特征表示能力，但是难以捕获图像的全局信息[12]。这就使基于卷积神经网络架构的图像语义编解码器难以兼顾图像的全局语义特征和局部语义特征。

2) 图像语义模型训练。基于单网络架构和传统损失函数的图像语义模型在上采样时容易丢失图像细节，当模型学习达到一定的精度后，容易陷入局部极值，使重建的图像语义信息失真[13]。

3) 图像语义评估指标。传统通信系统的评估指标误码率（BER, bit error ratio）并不适用于语义通信系统[1]。目前的图像语义通信模型一般针对某个特定的下游任务来评估语义通信的质量，无法对图像语义模型做出准确全面的评估，需要针对不同图像语义通信任务设计合适的评估指标。

针对以上问题，本文提出一种深度图像语义通信（DeepISC, deep image semantic communication）模型，主要的研究工作如下。

1) 设计了一种基于 vision transformer[14]的自编码器（ViTA, vision transformer-based autoencoder）网络来实现图像语义通信系统的建模。ViTA 引入 vision transformer 结构，首先使用 vision transformer encoder 和 vision transformer decoder 分别作为图像语义编码器和解码器。然后使用一个由深度神经网络（DNN, deep neural network）组成的全连接自编码器[15]作为信道编码器和解码器。由于引入了

vision transformer，ViTA 可以在发送端对传输图像进行更加准确的语义特征提取，在接收端可以根据接收到的语义特征进行更加精确的图像重建。由此，图像语义编解码器的能力得到提高。

2) 设计了一种用于指导 ViTA 学习的判别器网络（DSN, discriminator network），与 ViTA 形成双网络架构来进行协同训练，增强 ViTA 的学习能力。DSN 对 ViTA 生成的重建图像和原图像进行识别，根据识别的结果计算相应的对抗损失函数，并以此来指导 ViTA 的更新。另外，DSN 与 ViTA 保持同步更新，保证 DSN 始终具有对 ViTA 进行指导的能力。双网络架构和对抗损失函数可以更准确地捕捉人类对图像语义失真的感知，使重建图像在语义上更加接近原图像。由此，图像语义模型的训练效果得到改善。

3) 提出了一种针对下游任务的多任务图像语义评估机制。通过重新设计不同的图像语义评估指标，评估原图像和重建图像在 3 个下游视觉任务（图像分类、目标检测和特征提取）上的表现，从而更加准确地判断它们的语义相近程度。由此，图像的语义可以得到更精准的评估。

1 系统模型

如图 1 所示，本文考虑了一个发送端和接收端进行端到端通信的图像语义通信系统模型。该图像语义通信系统模型主要包括发送端、信道、接收端和图像语义评估 4 个部分。其中，发送端的功能主要包括对图像进行特征提取和语义编码、信道编码和信道调制；信道负责对信息进行传输；接收端的功能主要包括信道解调、信道解码和根据语义解码进行图像重建；图像语义评估则基于不同的下游任务以及传统的通信指标对图像语义通信的结果进行评估。

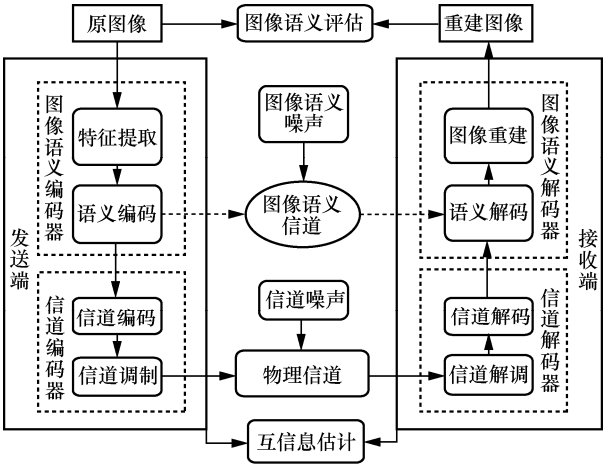

图 1 图像语义通信系统模型

1.1 问题描述

本文提出的图像语义通信的目标是在确定发射图像信号 $\boldsymbol{x}$ 的维度大小 M 的前提下，尽可能地使接收端重建的图像有效地还原图像的语义特征。该模型的设计主要面临 2 个困难，一是如何联合设计图像语义编码器和信道编码器；二是如何克服图像语义噪声，进行有效的语义传输，这是传统的通信系统没有考虑的问题。图像语义噪声是图像语义通信过程中引入的可能引起图像语义信息错误识别和解释的噪声，它可能在编码、传输和解码过程中产生。

1.2 图像语义编码器和解码器

如图 1 所示，发送端由图像语义编码器和信道编码器两部分组成。其中，图像语义编码器对输入的原图像 m 进行特征提取和语义编码，信道编码器负责信道编码和信道调制以保证编码后的语义信息在物理信道上能顺利传输。因此，发送端的发射信号可表示为

$$\boldsymbol{x} = C(S(m, \vartheta), \alpha) \tag{1}$$

其中，$\boldsymbol{x} \in \mathbb{R}^{M\times 1}$ 是一个 M 维的向量；$S(\cdot)$ 是图像语义编码器；ϑ 是图像语义编码器的参数集合；$C(\cdot)$ 是信道编码器；α 是信道编码器的参数集合。发送端将 $\boldsymbol{x}$ 发送出去，通过物理信道到达接收端，接收端得到的信号可表示为

$$\boldsymbol{y} = h\boldsymbol{x} + n \tag{2}$$

其中，$\boldsymbol{y} \in \mathbb{R}^{M\times 1}$ 是一个 M 维的向量；h 是瑞利衰落信道的信道增益；n 是加性白高斯噪声（AWGN, additive white Gaussian noise）。对于编码器和解码器的端到端训练，物理信道必须允许反向传播。因此，物理信道可以通过一个神经网络来进行模拟。

接收端包括图像语义解码器和信道解码器，分别用于恢复发送的信号和对图像进行语义重建。因此，经过解码后的重建图像可表示为

$$\hat{m} = S^{-1}(C^{-1}(\boldsymbol{y}, \beta), \delta) \tag{3}$$

其中，$\hat{m}$ 是重建图像；$C^{-1}(\cdot)$ 是信道解码器；β 是信道解码器的参数集合；$S^{-1}(\cdot)$ 是图像语义解码器；δ 是图像语义解码器的参数集合。本文的目标是尽可能地重建出与原图像相似的图像，因此可以使用均方误差（MSE, mean square error）作为图像语义模型的目标函数，即

$$\mathcal{L}_{\mathrm{MSE}}(m,\hat{m})=\min_{\vartheta,\ \delta}(m-\hat{m})^2 \tag{4}$$

通过最小化 MSE，图像语义网络可以学习原图像 m 中的像素分布，并进行图像重建，由此得到图像语义编码器和解码器的参数集合 ϑ 和 δ。

1.3　信道编码器和解码器

通信系统模型设计的一个重要目标是最大化通信系统容量或数据传输速率。互信息一般用来衡量 2 个变量之间的相关性，与误码率相比，互信息可以为训练接收机提供额外的信息[16]。为了提高系统容量，降低信道噪声对通信传输过程的影响，提升图像语义通信系统的稳健性，本文考虑最大化图像语义通信中信道输入和输出间的互信息。信道输入 $\boldsymbol{x}$ 和输出 $\boldsymbol{y}$ 之间的互信息可表示为

$$I(\boldsymbol{x};\boldsymbol{y})=\mathbb{E}_{p(x,y)}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right]=$$
$$\mathbb{E}_{p(x,y)}[\log p(\boldsymbol{y}\mid\boldsymbol{x})-\log p(\boldsymbol{y})] \tag{5}$$

其中，$(\boldsymbol{x},\boldsymbol{y})$ 为输入空间和输出空间里的随机变量对；$p(\boldsymbol{x})$ 为发射信号 $\boldsymbol{x}$ 的边缘概率分布；$p(\boldsymbol{y})$ 为接收信号 $\boldsymbol{y}$ 的边缘概率分布；$p(\boldsymbol{x},\boldsymbol{y})$ 为 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的联合概率分布；$p(\boldsymbol{y}\mid\boldsymbol{x})$ 为在给定 $\boldsymbol{x}$ 的条件下得到 $\boldsymbol{y}$ 的概率分布。

1.4　图像语义评估

合理的性能评估对图像语义系统的设计是非常重要的。在传统的端到端通信系统中，通常把 BER 作为评估的性能指标[17]，也就是主要考虑如何准确和有效地将符号或比特从发送端传送到接收端。然而与传统的端到端通信不同，图像语义通信旨在传送与目标图像语义相关的信息，会忽略许多语义无关的信息。因此 BER 并不适合作为图像语义通信系统的评估指标。

已有的图像语义通信模型通常基于某个特定的下游任务来对图像语义通信进行语义层面的评估。文献[16]基于图像分类任务来对图像语义通信的结果进行评估，将图像分类常用的损失函数交叉熵作为图像语义通信模型的评估指标。这种方式针对其他图像下游任务（如目标检测、行为跟踪等）往往不适用，亟须设计适合不同图像语义通信任务的评估指标。

2　深度图像语义通信系统

2.1　DeepISC 的设计与运行流程

如图 2 所示，DeepISC 模型主要由 ViTA、DSN 和图像语义评估 3 个模块组成。

1) 面向图像的语义通信网络 ViTA。ViTA 用于实现整个图像语义通信过程。具体来说，原图像输入 ViTA 中，依次经过图像语义编码器、信道编码器、物理信道、信道解码器和图像语义解码器后得到重建图像。ViTA 的目标是实现高效的图像语义特征信息提取、传输以及图像语义特征重建，使接收的图像拥有与原图像一致的语义。

2) 指导 ViTA 的判别器网络 DSN。DSN 对 ViTA 生成的重建图像和原图像进行识别，根据识别的结果计算相应的对抗损失函数，以此来指导 ViTA 的更新。具体来说，DSN 通过告诉 ViTA 它所生成的重建图像和原图像之间的语义差距，使 ViTA 可以朝着更好的方向进行学习。另外，DSN 与 ViTA 同步进行更新，也就是说 ViTA 生成的重建图像越逼

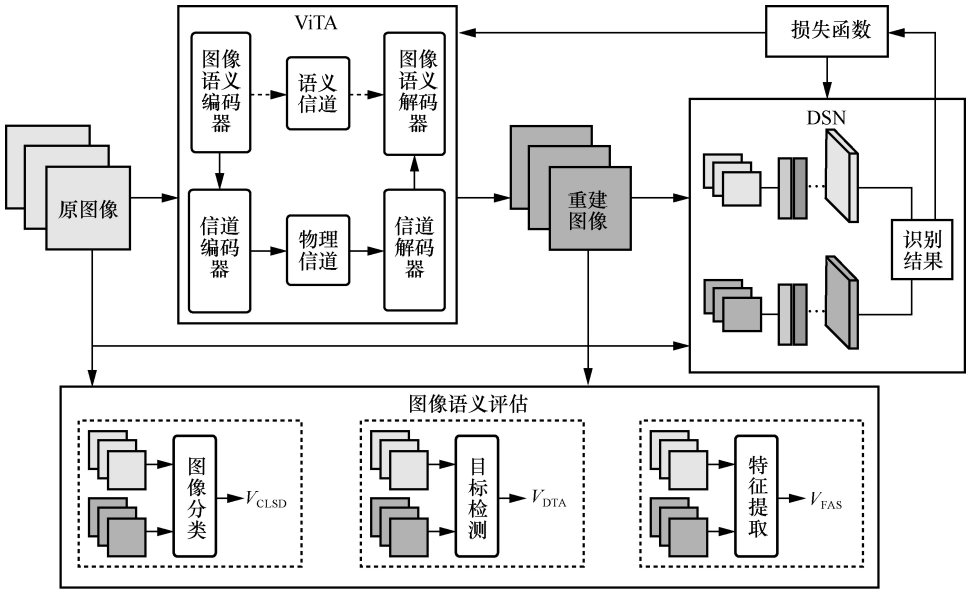

图 2　DeepISC 模型

真，DSN 的判别能力越强，因此 DSN 可以始终保持对 ViTA 的指导能力。

3) 图像语义评估。除了传统的通信评估指标外，本文还考虑将重建图像应用于不同的下游任务中，设计了多任务图像语义评估机制来评估重建图像与原图像的语义一致性。具体来说，在图像分类任务中，根据分类结果计算分类偏差值（CLSD, classification deviation）V_{CLSD}；在目标检测任务中，根据目标检测结果计算检测准确值（DTA, detection accuracy）V_{DTA}；对于特征提取的任务，根据特征向量之间的余弦距离计算特征相似值（FAS, feature similarity）V_{FAS}。最后根据原图像与重建图像在各个任务上的综合表现来评估语义信息的传输准确性。

2.2 面向图像语义的自编码器网络 ViTA

与传统的 CNN 相比，vision transformer 近年来在各种视觉任务（如图像分类、目标检测、特征提取等）中都展现出了更强的特征分析能力[18-19]。因此，为了更好地对图像进行特征提取和语义重建，本文提出的 ViTA 使用 vision transformer 作为 DeepISC 模型中的图像语义编码器和解码器部分，利用一个由多层 DNN 组成的自编码器作为信道编码器和解码器的部分，并使用一个感知器模型来模拟物理信道。

图 3 展示了 ViTA 的网络结构，在发送端，首先原图像经过 PatchEmbed 层，被调整成一系列一维的 Patch Embedding[20]，用 $\boldsymbol{E}$ 表示。接着图像语义编码器对 $\boldsymbol{E}$ 进行语义特征提取，得到原图像 $\boldsymbol{m}$ 的语义编码。然后使用信道编码器对语义编码进行信道编码，根据式(1)可知，最后发送端的发射信号为 $\boldsymbol{x}$。其中图像语义编码器主要由 vision transformer 编码器构成，vision transformer 编码器的核心是多头注意力层。多头注意力层中的多头注意力机制可以更好地学习某一像素点和其他位置，包括较远位置的像素点的关系，这样可以更好地获取图像的特征信息并使重建图像效果更加逼真[21]。多头注意力层本质上是由多个自注意力头 head_i 拼接组成的，一个自注意力头由一个自注意力层得到，计算式为

$$\mathrm{head}_i = \mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (6)$$

其中，$\boldsymbol{Q}$ 为查询向量；$\boldsymbol{K}$ 是与 $\boldsymbol{Q}$ 相匹配的匹配向量；$\boldsymbol{V}$ 为信息向量[22]；$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 均由输入 $\boldsymbol{E}$ 进行线性变换得到，即 $\boldsymbol{Q}=\boldsymbol{E}\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{K}=\boldsymbol{E}\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{V}=\boldsymbol{E}\boldsymbol{W}_{\boldsymbol{V}}$，$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}$ 为各自的权重向量；d_k 为调节因子。

多个自注意力头进行拼接即可得到多头注意力，计算式为

$$\mathrm{MultiheadAttention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right) = \mathrm{concat}\left(\mathrm{head}_1,\cdots,\mathrm{head}_h\right)\boldsymbol{W}_{\mathrm{mha}} \quad (7)$$

其中，h 表示拼接的自注意力头的数量；$\boldsymbol{W}_{\mathrm{mha}}$ 表示多头注意层的权重；concat(·) 表示向量的拼接运算。信道编码器由两层隐藏单元数不同的 DNN 组成，隐藏单元数逐层递减以进行数据压缩。第一层的激活函数采用 ReLU 函数，第二层的激活函数采用 tanh 函数，编码后的输出 $\boldsymbol{x}$ 可表示为

$$\boldsymbol{x} = \tanh\left(\mathrm{ReLU}\left(S(m,\vartheta)\boldsymbol{W}_1+\boldsymbol{b}_1\right)\boldsymbol{W}_2+\boldsymbol{b}_2\right) \quad (8)$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为 DNN 的权重矩阵，$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 为 DNN 的偏置。

为了实现信道上编码器与解码器的联合训练，本文使用一层感知器神经网络来模拟信道，该神经网络本质上是一种输入和输出的映射关系，主要由该层网络的神经元权重 $\boldsymbol{W}_n$ 和偏置 $\boldsymbol{b}_n$ 决定。其中，权重 $\boldsymbol{W}_n$ 是信道增益；偏置 $\boldsymbol{b}_n$ 是为了模拟信道中的加性白高斯噪声 $\boldsymbol{n}$ 而添加的高斯随机变量，该变量的方差对应 AWGN 的功率，它的值主要由信噪比以及发送功率决定。信号 $\boldsymbol{x}$ 在物理信道中进行传播，受到信道噪声的干扰后，根据式(2)，到达接收端时变成信号 $\boldsymbol{y}$。

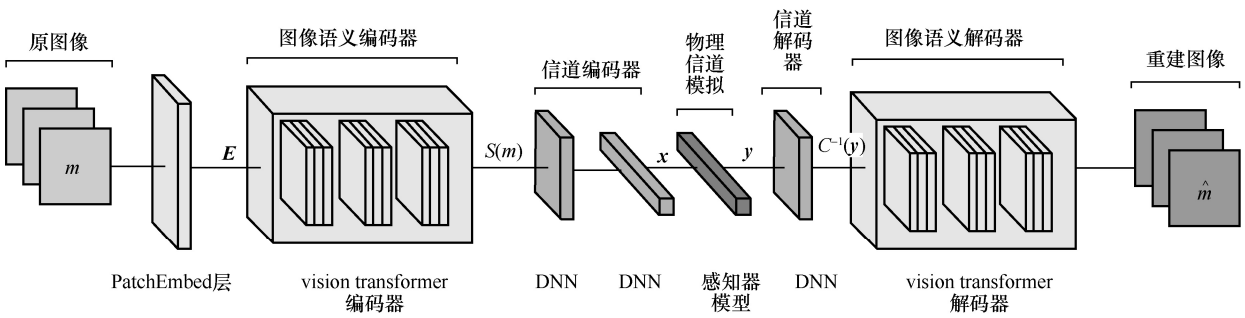

图 3 ViTA 的网络结构

在接收端，信道解码器由一层 DNN 构成，负责对信号 $\boldsymbol{y}$ 进行信道解码；图像语义解码器主要由 vision transformer 解码器构成，负责利用解码后的特征信息进行图像重建，最后得到 $\hat{m}$ 。

2.3　判别器网络 DSN

已有的图像语义通信模型大多基于单网络和传统的损失函数进行训练，这样的训练方式使模型存在陷入局部最优的风险。为此，本文提出一种使用 DSN 和 ViTA 组成双网络架构进行协同训练的方式。接下来，本节将依次介绍 DSN 的网络结构、ViTA 和 DSN 的对抗损失函数的定义以及它们协同训练的过程。

如图 4 所示，DSN 的网络结构由特征提取层和输出层组成，其中特征提取层由 4 层 CNN 和 4 层 InstanceNorm 层组成。InstanceNorm 在一个通道内做归一化，计算特征图的高×宽的均值，这样的方式可以加速模型收敛，并且保持每个图像实例之间的独立性[23]，其计算式为

$$\boldsymbol{y}=\frac{\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]}{\sqrt{\mathrm{Var}[\boldsymbol{x}]+\epsilon}}\gamma+\omega \tag{9}$$

其中，$\boldsymbol{x}$ 和 $\boldsymbol{y}$ 表示输入和输出；$\mathrm{Var}[\cdot]$ 表示方差运算；γ 、ω 、ϵ 表示调节参数。图像输入 DSN 后，首先由 CNN 层和 InstanceNorm 层进行特征提取，然后到输出层进行决策。

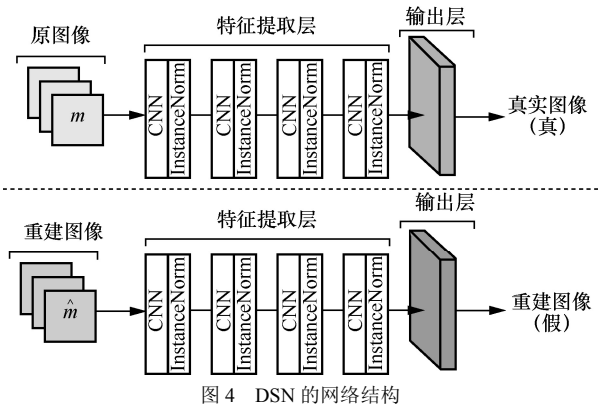

图 4　DSN 的网络结构

传统的图像语义网络模型都是仅基于式(4)来更新整个语义网络模型的，这样存在的问题是图像语义模型在上采样时容易丢失图像细节，导致重建图像语义信息失真。因此本文提出 DSN 和 ViTA 协同训练以解决上述问题。假设 G 表示 ViTA 模型，则 ViTA 和 DSN 之间的对抗损失函数可以表示为

$$\mathcal{L}_{\text{against}}=\mathbb{E}_{m}[\log D(m)]+\mathbb{E}_{m,\boldsymbol{W}_G}[\log(1-D(G(m;\boldsymbol{W}_G)))] \tag{10}$$

其中，$D(\cdot)$ 表示 DSN 模型的输出，输出值的范围为 $[0,1]$；$\boldsymbol{W}_G$ 表示模型 G 的参数向量；$G(m;\boldsymbol{W}_G)$ 表示图像 m 输入模型 G 后得到的结果，在本文中为 $\hat{m}$ 。因此，式(15)可以表示为

$$\mathcal{L}_{\text{against}}=\mathbb{E}_{m}[\log D(m)]+\mathbb{E}_{\hat{m}}[\log(1-D(\hat{m}))] \tag{11}$$

在 DSN 和 ViTA 的协同训练中，DSN 的目标是能够分辨出原图像 m 和 ViTA 生成的重建图像 $\hat{m}$ ，具体来说就是将 m 识别为真，将 $\hat{m}$ 识别为假。而 ViTA 则是努力生成逼真的重建图像 $\hat{m}$ 以混淆 DSN 的判断。因此，ViTA 和 DSN 协同训练的总目标可以表示为

$$\mathcal{L}_{\text{total}}=\arg\min_{G}\max_{D}\mathcal{L}_{\text{against}} \tag{12}$$

为了使 ViTA 模型更快地收敛，参考文献[24]的工作与结论，本文在对抗损失函数中加入 L1 损失，L1 损失可表示为

$$\mathcal{L}_{\text{L1}}=\mathbb{E}_{m,\hat{m}}[|\hat{m}-m|] \tag{13}$$

其中，$|\cdot|$ 表示绝对值运算。所以指导 ViTA 更新的损失函数可表示为

$$\mathcal{L}_{\text{AiT}}=\mathcal{L}_{\text{against}}+\lambda\mathcal{L}_{\text{L1}} \tag{14}$$

其中，λ 是调节 L_1 权重大小的系数。接下来，介绍用于更新 DSN 的损失函数，可表示为

$$\mathcal{L}_{D}=-\left[y_m\ln\left(\frac{1}{1+\mathrm{e}^{-D(x_m)}}\right)+(1-y_m)\ln\left(1-\frac{1}{1+\mathrm{e}^{-D(x_m)}}\right)\right] \tag{15}$$

其中，x_m 代表输入的图像；y_m 代表该图像的标签值，如果 x_m 是原图像，则 $y_m=1$；如果输入是重建图像，即 $x_m=x_{\hat{m}}$，则其对应的标签 $y_m=0$ 。

本文训练的方式是先训练 ViTA 中的信道编码器和解码器，待其收敛后将其参数固定，然后用对抗的方式对 ViTA 和 DSN 进行协同训练。

2.4　多任务图像语义评估机制

为了更好地对图像语义通信的结果进行语义层级上的评估，本文提出一种针对下游任务的多任务图像语义评估机制。具体来说，本文通过重新设计不同的图像语义评估指标，评估原图像和重建图像在 3 个下游视觉任务上的表现，判断它们的语义相似度。

如图 5 所示，本文考虑 3 个常见的视觉任务，即图像分类、目标检测和特征提取，并提出以下 3 个图像语义的评估指标。

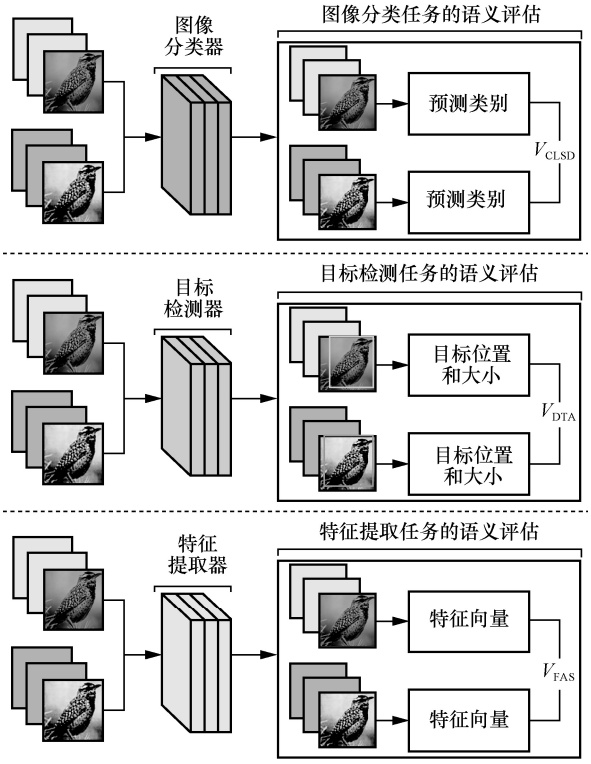

图 5　图像语义评估

1) 分类偏差值。该指标主要针对图像分类任务。利用图像分类器对原图像和重建图像分别进行识别，然后根据分类器的识别结果计算得出分类偏差值。假设总共有 N 个类别，分类偏差值可表示为

$$V_{\mathrm{CLSD}}=\frac{1}{N}\sum_{i=1}^{N}(C_i(\hat{m})-C_i(m))^2 \tag{16}$$

其中，$C_i(m)$ 和 $C_i(\hat{m})$ 分别代表原图像 m 和重建图像 $\hat{m}$ 被识别为第 i 个类别的概率。因此，V_{CLSD} 越小说明原图像 m 和重建图像 $\hat{m}$ 在分类任务上的语义越接近，反之则说明两者的语义相差越大。

2) 检测准确值。该指标主要针对图像中的目标检测任务。使用目标检测器分别对原图像和重建图像进行目标检测，根据检测结果计算得出两者的检测准确值。假设目标检测器对原图像识别出的目标框集合为 $\mathfrak{R}_1$，对重建图像识别出的目标框集合为 $\mathfrak{R}_2$，检测准确值可以表示为

$$V_{\mathrm{DTA}}=\sum_{i\in\mathfrak{R}_1}\xi\max(\{U(i,j)\mid\forall j\in\mathfrak{R}_2\}) \tag{17}$$

其中，ξ 表示框 i 和框 j 的预测类别是否相同，如果相同，则 $\xi=1$，否则 $\xi=0$；$U(i,j)$ 表示框 i 和框 j 之间的交并比（IoU）[25]，其计算式为

$$U(i,j)=\frac{O(i,j)}{S(i)+S(j)-O(i,j)} \tag{18}$$

其中，$O(i,j)$ 表示框 i 和框 j 的重合部分的面积；$S(i)$ 和 $S(j)$ 分别代表框 i 和框 j 的面积大小，当框 i 和框 j 完全重合时 $U(i,j)=1$，当框 i 和框 j 完全分离时 $U(i,j)=0$。因此，V_{DTA} 越大说明原图像 m 和重建图像 $\hat{m}$ 在目标检测任务上的语义越接近，反之则说明两者的语义相差越大。

3) 特征相似值。该指标主要检测重建图像在特征提取任务上是否可以保持和原图像相似的语义。首先使用一个特征提取器对原图像和重建图像分别进行特征提取，然后计算两者的特征距离。本文主要针对图像数据，因此采用余弦距离来衡量特征之间的距离。由此，特征相似值可表示为

$$V_{\mathrm{FAS}}=\cos(f(m),f(\hat{m})) \tag{19}$$

其中，$\cos(\cdot)$ 表示余弦距离；$f(m)$ 和 $f(\hat{m})$ 分别表示对原图像和重建图像进行特征提取后的结果。因此，V_{FAS} 越大说明两者的特征相似度越高，即语义越相近；反之则说明重建图像不能在该任务上很好地表达原图像的语义。

以上 3 个针对不同下游视觉任务的图像语义评估指标构成了本文所提出的多任务图像语义评估机制。

3　仿真分析

3.1　实验设置

本文实验采用的数据集为来自 Kaggle 的公开数据集 BIRDS-400[26]。实验的训练和测试环境为 Ubuntu 20.04+CUDA 10.2，编程语言为 Python，使用的深度学习框架为 PyTorch1.8.0。下面分别介绍 3 种下游测试任务和基于传统通信指标峰值信噪比（PSNR, peak signal to noise ratio）的实验设置。

1) 图像分类任务。本文实验使用 3 种常用网络 ResNet50、VGG16 和 AlexNet[27]作为分类器网络。然后在训练集上对 3 个网络分别进行训练得到 3 个对应的公共图像分类器。

2) 目标检测任务。本文实验采用 YOLOv4[28]作为公共检测器，由于 YOLOv4 的预训练模型已经可以较好地对鸟类图片进行目标检测，因此不需要训练就可直接使用。

3) 特征提取任务。本文实验同样采用 ResNet50、VGG16 和 AlexNet 作为特征提取网络。3 个网络都先

加载各自的预训练权重，并只保留特征层的部分；本文采用余弦距离来衡量特征之间的差距。

4) 基于 PSNR 的语义通信评估。本文实验采用 PSNR 对原图像和基于不同通信模型生成的重建图像进行评估，判断语义通信中图像的失真程度。

本文考虑常用的变分自编码器（VAE, variational autoencoder）、卷积自编码器（CAE, convolutional autoencoder）[29]以及无 DSN 指导的 ViTA（VAwD, ViTA without DSN）作为基准模型。本文实验的所有模型的训练回合数都设置为 400，数据批次大小为 32，学习率为 0.0001，采用 Adam 优化器[30]。

3.2　实验结果分析

3.2.1　收敛性实验

为评估本文所提的 DeepISC 模型的收敛性，本节进行了收敛性分析实验，实验结果如图 6 所示。

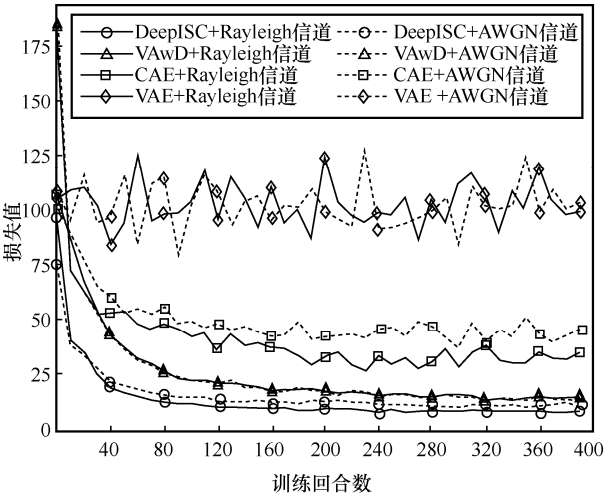

图 6　不同语义通信模型的损失值随训练回合数的变化

从图 6 可以看出，在 Rayleigh 和 AWNG 这 2 种信道上，4 种通信模型在经过一定轮次的训练后最终都达到了收敛。其中，DeepISC 达到了最优的收敛值，VAwD 次之，VAE 和 CAE 模型的效果都明显低于前面 2 种。

图 6 的实验结果表明，本文所提的 DeepISC 经过有限次的训练后可以达到收敛，并且相比于其他几种模型，可以达到更好的收敛结果。

3.2.2　基于图像分类任务的语义评估

为评估 DeepISC 模型在图像分类任务上的语义表现，本节的实验对比了 DeepISC 和 VAE、CAE 以及 VAwD 模型在 AWGN 和 Rayleigh 信道上的表现。为了方便观察，这里取分类偏差值的对数作为实验指标，实验结果如图 7 所示。

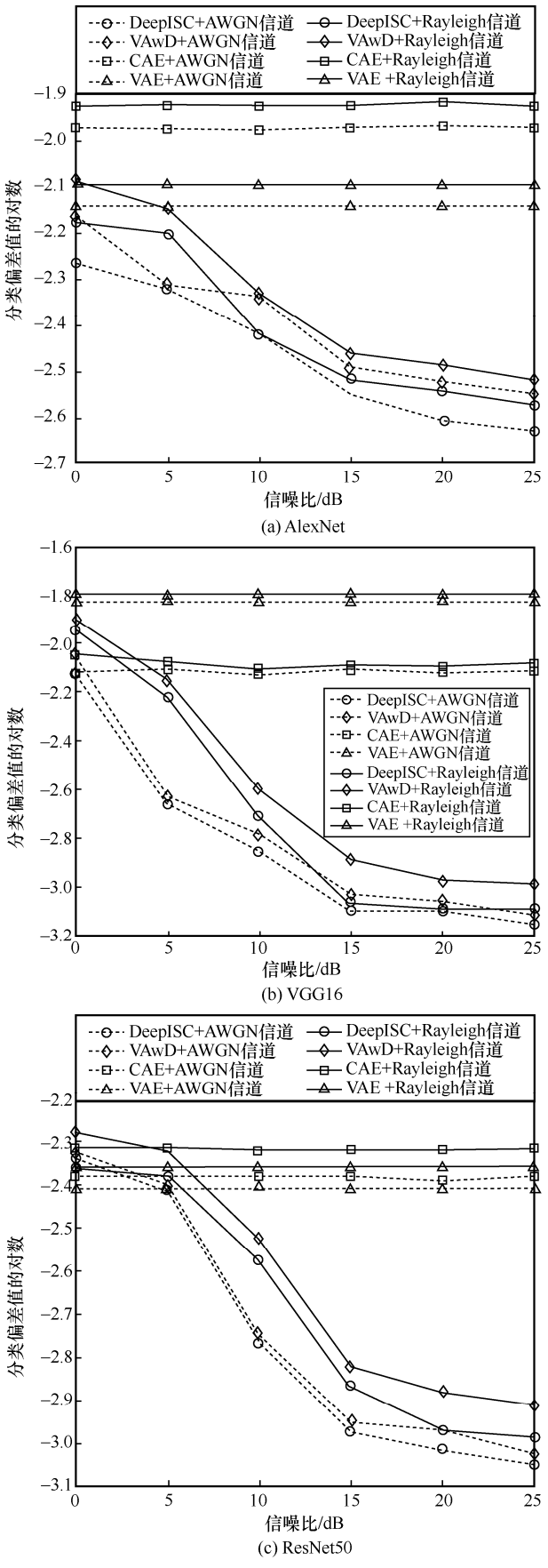

图 7　不同语义通信模型的分类偏差值随信噪比的变化

从图 7(a)可以看出，当 AlexNet 作为分类器时，无论是在 AWGN 信道还是在 Rayleigh 信道上，DeepISC 的表现都是最好的，VAwD 的表现次之，VAE 和 CAE 较差。从图 7(b)和图 7(c)可以看出，当 VGG16 或者 ResNet50 作为分类器时，在 AWGN 和 Rayleigh 信道上，除了在低 SNR 的情况之外，DeepISC 的效果都是最好的，VAwD 次之，CAE 和 VAE 较差。而且 DeepISC 只有在 ResNet50 作为分类器且信道为 Rayleigh 时，才在低 SNR 下表现出了较低的性能。

图 7 的实验结果表明本文所提出的 DeepISC 在 2 种不同的信道下，相较于其他基准模型在图像分类任务上有总体更好的语义表现，并且在 SNR 有所改善时，性能提升非常明显。另外，当 SNR 较高时，DeepISC 在 AWGN 和 Rayleigh 信道上的表现差距并不大，说明在面对图像分类任务时，DeepISC 对不同信道的泛化能力较强。

3.2.3 基于目标检测任务的语义评估

为了验证所提出的 DeepISC 模型相较于其他的基准模型在目标检测的任务上有更好的语义表现，本节实验对比了 DeepISC、VAE、CAE 以及 VAwD 在 AWGN 和 Rayleigh 信道上的表现，验证了在不同信噪比条件下，YOLOv4 作为目标检测器时检测准确值 V_{DTA} 的变化，实验结果如图 8 所示。

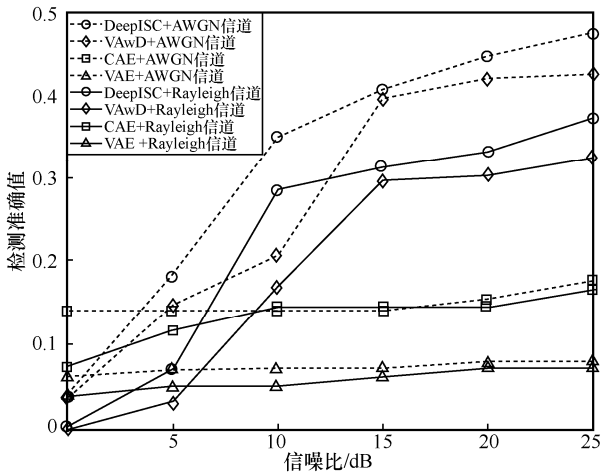

图 8 不同语义通信模型的检测准确值随信噪比的变化

从图 8 可以看出，在 AWGN 信道上，当 SNR=0 时，VAE 和 CAE 效果较优，而当 SNR>0 时，DeepISC 和 VAwD 的表现有明显的提升。随着 SNR 的提高，DeepISC 的检测准确值也在逐步提高。在 Rayleigh 信道上，当 0<SNR<5 时，CAE 的效果最优，但随着 SNR 的改善，DeepISC 和 VAwD 的检测准确值很快超过了 CAE 和 VAE。

图 8 的实验结果表明，本文所提出的 DeepISC 在 2 种信道环境下，当 SNR 较好时，相较于其他基准模型在目标检测的任务上都有更好的语义表现，并且在 SNR 有所改善时，性能提升非常明显。

3.2.4 基于特征提取任务的语义评估

为评估所提出的 DeepISC 模型相较于其他的基准模型在特征提取任务上有更好的语义表现，本节实验测试了在不同信道信噪比条件下，4 种模型在 3 个特征提取网络上的特征相似值 V_{FAS} 的变化，实验结果如图 9 所示。

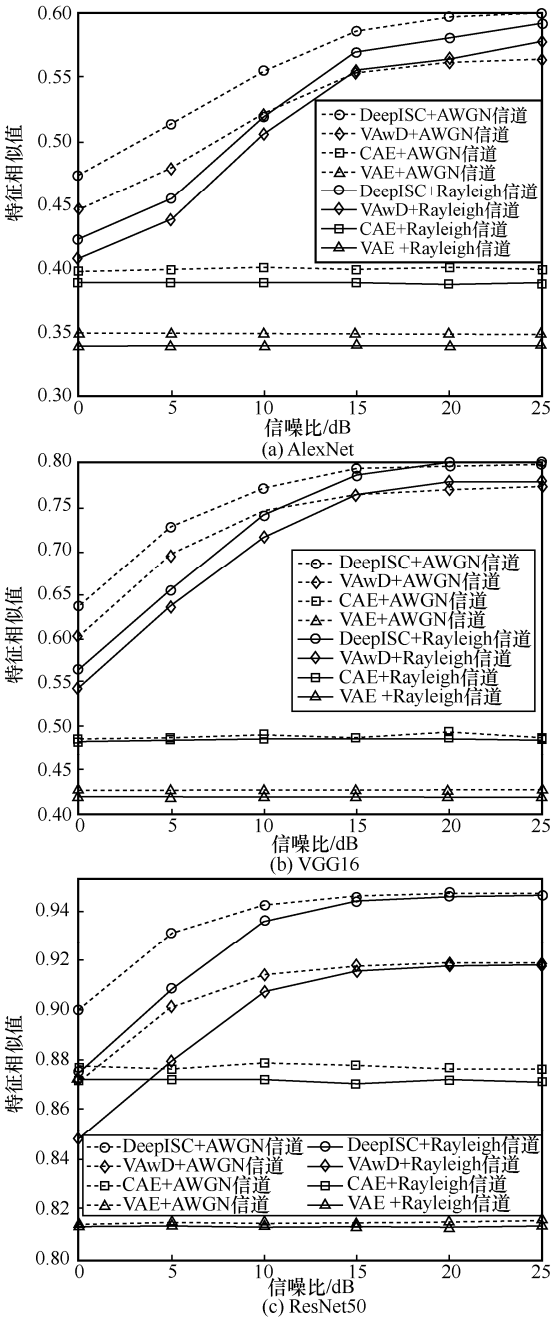

图 9 不同语义通信模型的特征相似值随信噪比的变化

从图 9(a)和图 9(b)可以看出，当 AlexNet 和 VGG16 作为特征网络时，在 AWGN 信道和 Rayleigh 信道上，本文所提出的 DeepISC 的性能都是最好的，VAwD 的性能仅次于 DeepISC，而 VAE 和 CAE 性能较差。从图 9(c)可以看出，当 ResNet50 作为特征网络时，在 AWGN 信道上，DeepISC 的性能仍然是最好的，仅当 SNR=0 时，CAE 的性能优于 VAwD；在 Rayleigh 信道上，当 SNR=0 时，DeepISC 的性能仅次于 CAE，但随着 SNR 的提高，DeepISC 和 VAwD 的性能逐渐超过了 CAE，并且 DeepISC 的性能一直保持最优。

图 9 的实验结果表明，本文所提出的 DeepISC 在 2 种信道下的性能都达到了最优，说明 DeepISC 在基于特征提取任务上的语义通信可以达到一个较优的结果。另外 DeepISC 在 2 种信道上的表现差距不大，说明在面对特征提取任务时，DeepISC 对不同信道的适应能力较强。

3.2.5　基于 PSNR 的语义通信评估

本节采用 PSNR 作为评估指标，以评估在语义通信中的图像失真程度，实验结果如图 10 所示。

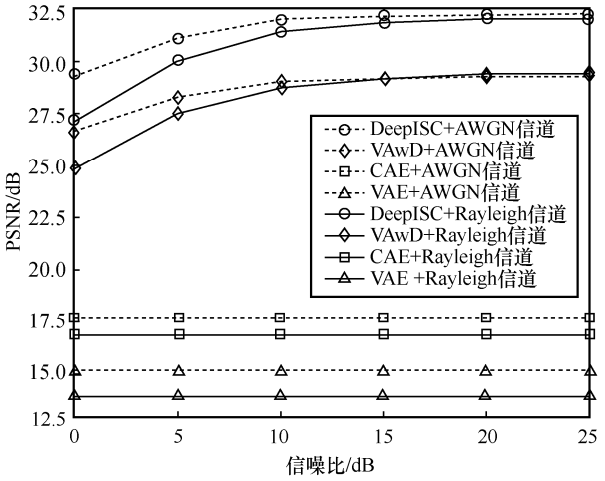

图 10　不同语义通信模型的峰值信噪比随信噪比的变化

从图 10 可以看出，4 种模型在 2 种信道上的 PSNR 的实验结果基本相同，都是 DeepISC 最优，VAwD 次之，CAE 较差，VAE 最差。这说明相比于其他模型，DeepISC 模型重建的图像失真程度最低，保持了与原图像在像素层面的一致性。

图 10 的实验结果表明，本文提出的 DeepISC 在传统的评估指标上同样有较好的表现，进一步证明了由 DeepISC 模型实现语义通信的有效性。

4　结束语

目前的图像语义通信模型在处理图像数据时存在图像语义编码器对图像语义特征信息提取不够准确，接收端对重建的图像语义信息不够准确以及缺乏合理的面向图像语义通信任务的评估指标等问题。针对以上问题，本文提出了 DeepISC 模型，该模型首先采用一种基于 vision transformer 构造的 ViTA 网络来实现图像语义编解码；其次利用 DSN 和 ViTA 的双网络架构来进行协同训练，提高重建图像的语义精度；然后设计了一种通用的多任务图像语义评估机制，能够对多种常见的视觉语义通信任务进行全面评估，拓展了 DeepISC 的应用场景；最后仿真结果验证了本文提出的 DeepISC 模型的可行性。

未来工作将研究如何改善该图像语义通信模型在低信噪比下的性能表现以及在更复杂信道中的泛化能力，使其可以广泛地应用。另外，笔者将考虑如何对图像语义通信模型进行有效压缩，减少发送端和接收端的计算开销，进一步提升通信性能的同时降低时延，从而保证语义通信系统的实时性。

参考文献：

[1] 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.

[2] TSE D, VISWANATH P. Fundamentals of wireless communication[M]. Cambridge: Cambridge University Press, 2005.

[3] XIE H Q, QIN Z J. A lite distributed semantic communication system for Internet of things[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(1): 142-153.

[4] TONG H N, YANG Z H, WANG S H, et al. Federated learning based audio semantic communication over wireless networks[C]//Proceedings of 2021 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2021: doi.org/10.1109/GLOBECOM46510.2021.9685654.

[5] KOTSAKIS R, KALLIRIS G, DIMOULAS C. Investigation of broadcast-audio semantic analysis scenarios employing radio-programme-adaptive pattern classification[J]. Speech Communication, 2012, 54(6): 743-762.

[6] HUANG D L, TAO X M, GAO F F, et al. Deep learning-based image semantic coding for semantic communications[C]//Proceedings of 2021 IEEE Global Communications Conference (GLOBECOM). Piscataway: IEEE Press, 2021: doi.org/10.1109/GLOBECOM46510.2021: 9685667.

[7] PATWA N, AHUJA N, SOMAYAZULU S, et al. Semantic-preserving image compression[C]//Proceedings of 2020 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2020:

1281-1285.

[8] SUN Q Z, GUO C L, YANG Y, et al. Deep joint source-channel coding based on semantics of pixels[J]. arXiv Preprint, arXiv: 220811375, 2022.

[9] WANG J, WANG S X, DAI J C, et al. Perceptual learned source-channel coding for high-fidelity image semantic transmission[C]//Proceedings of IEEE Global Communications Conference. Piscataway: IEEE Press, 2023: 3959-3964.

[10] WANG Q, SHEN L Q, SHI Y. Recognition-driven compressed image generation using semantic-prior information[J]. IEEE Signal Processing Letters, 2020, 27: 1150-1154.

[11] HU Q Y, ZHANG G Y, QIN Z J, et al. Robust semantic communications against semantic noise[C]//Proceedings of 2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall). Piscataway: IEEE Press, 2022: doi.org/10.1109/VTC2022-Fall57202.2022.10012843.

[12] LIU X Y, WU Y, LIANG W K, et al. High resolution SAR image classification using global-local network structure based on vision transformer and CNN[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 1-5.

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

[14] PARMAR N, VASWANI A, USZKOREIT J, et al. Image transformer[C]//Proceedings of the 35th International Conference on Machine Learning. New York: PMLR, 2018: 4055-4064.

[15] PU Y C, GAN Z, HENAO R, et al. Variational autoencoder for deep learning of images, labels and captions[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM Press, 2016: 2360-2368.

[16] ZHANG H W, SHAO S, TAO M X, et al. Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 170-185.

[17] SHANNON C E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379-423.

[18] CHEN C F R, FAN Q F, PANDA R. CrossViT: cross-attention multi-scale vision transformer for image classification[C]//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2022: 347-356.

[19] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection[J]. arXiv Preprint, arXiv: 201004159, 2020.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv Preprint, arXiv: 201011929, 2020.

[21] ZHAO H S, JIA J Y, KOLTUN V. Exploring self-attention for image recognition[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 10073-10082.

[22] 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.

[23] KOHL S, BONEKAMP D, SCHLEMMER H P, et al. Adversarial networks for the detection of aggressive prostate cancer[J]. arXiv Preprint, arXiv: 170208014, 2017.

[24] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 5967-5976.

[25] YU J H, JIANG Y N, WANG Z Y, et al. UnitBox: an advanced object detection network[C]//Proceedings of the 24th ACM International Conference on Multimedia. New York: ACM Press, 2016: 516-520.

[26] WU W B, PAN Y. Adaptive modular convolutional neural network for image recognition[J]. Sensors, 2022, 22(15): 5488.

[27] THECKEDATH D, SEDAMKAR R R. Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks[J]. SN Computer Science, 2020, 1(2): 79.

[28] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv Preprint, arXiv: 200410934, 2020.

[29] PARIKH H, PATEL S, PATEL V. Evaluation of deep learning and transform domain feature extraction techniques for land cover classification: balancing through augmentation[J]. Environmental Science and Pollution Research, 2023, 30(6): 14464-14483.

[30] ZHANG Z J. Improved Adam optimizer for deep neural networks[C]//Proceedings of 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Piscataway: IEEE Press, 2019: 1-2.

[作者简介]

江沸菠（1982− ），男，湖南株洲人，博士，湖南师范大学副教授、硕士生导师，主要研究方向为深度学习与物联网等。

彭于波（1996− ），男，重庆人，湖南师范大学硕士生，主要研究方向为语义通信和联邦学习。

董莉（1982− ），女，湖南长沙人，博士，湖南工商大学讲师、硕士生导师，主要研究方向为深度学习与物联网等。