第48卷 第5期 计 算 机 学 报 Vol. 48 No. 5
2025年5月 CHINESE JOURNAL OF COMPUTERS May 2025

基于对比学习和扩散模型的图像语义通信方法

邢焕来[1),2),3)] 况田泽宇[1)] 徐乐西[4)] 李 洋[1),3)] 郑丹阳[1),3)]
罗寿西[1),3)] 戴朋林[1),2),3)] 李 可[1),3)] 冯 力[1),2),3)]

[1)](西南交通大学计算机与人工智能学院 成都 611756)
[2)](西南交通大学唐山研究院 河北 唐山 063000)
[3)](可持续城市交通智能化教育部工程研究中心 成都 611756)
[4)](中国联通研究院 北京 100048)

摘 要 无监督图像传输语义通信方法的目的是在接收端生成与原始图像尽可能相似的图像。然而,在低压缩比(Compression Ratios,CRs)或低信噪比(Signal-to-Noise Ratio,SNR)的条件下,重建图像的质量往往会显著下降。为了解决上述问题,本文提出了一种能够自监督提取并利用图像高层语义的语义通信模型。为有效地提取包含高、低层的混合语义特征,本文提出了一种基于动量对比(Momentum Contrast,MoCo)的自监督语义编码器。与纯高层语义特征相比,混合语义特征能够提高生成图像与原图的相似性;与纯低层语义特征相比,混合语义特征能够更好地表示图像的核心语义。本文设计了一种基于扩散模型(Diffusion Model)的语义解码器,旨在建立高层语义特征到图像像素的映射关系,将接收到的语义信息还原为图像像素。实验结果表明,在低CRs和低SNR条件下,与传统通信方法和四种无监督语义通信方法相比,本文所提出的模型在加性高斯白噪声(Additive White Gaussian Noise,AWGN)信道、瑞利(Rayleigh)信道和莱斯(Rician)信道下生成的图像在视觉特征方面与原数据集更接近,并在使用第三方图像分类器进行分类时具有更高的正确率。

关键词 语义通信;深度学习;对比学习;扩散模型;图像处理
中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.01151

Image Semantic Communication Method Based on Contrast Learning and Diffusion Model

XING Huan-Lai[1),2),3)] KUANG Tian-Ze-Yu[1)] XU Le-Xi[4)] LI Yang[1),3)] ZHENG Dan-Yang[1),3)]
LUO Shou-Xi[1),3)] DAI Peng-Lin[1),2),3)] LI Ke[1),3)] FENG Li[1),2),3)]

[1)](School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756)
[2)](Tangshan Institute, Southwest Jiaotong University, Tangshan, Hebei 063000)
[3)](Engineering Research Center of Sustainable Urban Intelligent Transportation, Ministry of Education, Chengdu 611756)
[4)](Research Institute, China United Network Communications Corporation, Beijing 100048)

Abstract In recent years, traditional communication methods have approached the Shannon limit. Meanwhile, deep learning has achieved revolutionary breakthroughs in areas such as large language models, computer vision and speech recognition. To address the challenge of exponentially

收稿日期:2024-06-26;在线发布日期:2025-03-14。本课题得到国家自然科学基金面上项目(No.62172342)、国家自然科学基金青年基金(No.62202392)、河北省自然科学基金(No.F2022105027)、中央高校基本科研业务费资助。**邢焕来(通信作者)**,博士,副教授,博士生导师,主要研究领域为语义通信、网络功能虚拟化、软件定义网络、人工智能与进化计算。E-mail:hxx@swjtu.edu.cn。**况田泽宇**,硕士研究生,主要研究领域为语义通信、深度学习。**徐乐西**,博士,教授级高工,主要研究领域为大数据、网络分析和智能运营。**李 洋**,博士,助理教授,主要研究领域为语义通信和边缘智能。**郑丹阳**,博士,特聘副研究员,硕士生导师,主要研究领域为网络功能虚拟化、在网计算、网络可靠性及安全性、数字孪生系统。**罗寿西**,博士,副教授,硕士生导师,主要研究领域为数据中心网络和网络化系统。**戴朋林**,博士,副教授,博士生导师,主要研究领域为智能交通系统和车辆信息物理系统。**李 可**,博士,副教授,硕士生导师,主要研究领域为车联网。**冯 力**,博士,研究员,博士生导师,主要研究领域为网络空间安全与人工智能。

growing network traffic, deep-learning-empowered semantic communication has emerged as a promising solution. Different from traditional communication methods focusing on the accuracy of symbol-level transmission, semantic communication focuses on the transmission of semantic information. This difference makes semantic communication noise-resistant and does not suffer from a cliff drop in performance in low signal-to-noise ratio (SNR) conditions. Compared to traditional communication methods, the semantic communication model allows to deliver identical semantic information with lower compression ratios (CRs) by sharing a semantic knowledge base between the transmitter and the receiver. So far, semantic communication has been studied in the context of text, speech, image, video and mixture of them. For image transmission, the mainstream semantic communication methods are unsupervised and supervised in terms of whether a method uses information other than images. Among them, unsupervised semantic communication methods for image transmission aim to generate an image as similar as possible to the original one at the receiver. However, the quality of the reconstructed image tends to significantly degrade under low CRs or low SNR. Some researchers have tried to solve this problem by introducing additional information to turn them into supervised methods. These methods mostly use semantic segmentation graphs or image labeling information as supervisory information and have a stronger semantic transfer capability compared to unsupervised methods. However, such a method relies on a large amount of manually labeled data or pre-trained models based on labeled data, which cannot automatically extract high-level semantic information from original images. To address the issue above, this paper proposes a semantic communication model that can self-supervise to extract and utilize the high-level semantics of an image, and in order to efficiently extract the hybrid semantic features containing both high— and low-levels, this paper proposes a self-supervised semantic coder based on Momentum Contrast (MoCo). Compared to purely high-level semantic features, mixed semantic features can improve the similarity between generated images and the original ones; compared to purely low-level semantic features, mixed semantic features can better represent the core semantics of images. This paper designs a semantic decoder based on the diffusion model, aiming to establish a mapping relationship between high-level semantic features and image pixels and restore the received semantic information to image pixels. The experimental results show that under low CRs and low SNR conditions, compared with the traditional communication methods and four unsupervised semantic communication methods, the images generated by the proposed model in this paper under Additive White Gaussian Noise (AWGN) channel, Rayleigh channel and Rician channel are closer to the original dataset in terms of visual features and have higher correctness rates when classified using third-party image classifiers. Furthermore, the proposed model demonstrates that contrastive learning can provide new sources of high-level semantics for semantic communication. In particular, analysis of high-level-to-low-level semantic ratios confirms the better noise immunity of high-level semantics compared to their low-level counterparts.

Keywords semantic communication; deep learning; contrastive learning; diffusion models; image processing

1 引 言

随着 5G 网络技术的广泛部署及其在各垂直领域的快速落地,网络流量将持续增长。然而,传统无线通信所支持的数据传输速率接近了香农极限[1]。语义通信是解决这一挑战的潜在技术。与传统通信相比,语义通信优先考虑语义信息的正确传递,而不

是符号的准确传递[2]。这种区别使得语义通信消耗更少的带宽资源,从而提高了数据传输效率。现有的语义通信研究针对多种数据类型,主要包括文本[3]、语音[4]、图像[5]和视频[6],且语义通信已经表现出优于传统通信方法的性能。

从图像传输和处理的角度来看,语义通信的研究大致分为两类:面向任务[7]和面向重建[8]。这两类研究均需从输入数据中提取和编码语义信息,但对语义信息和背景知识的需求有所区别[9]。

面向任务的语义通信系统通常优先传达与任务目标相关的细节,这些系统利用接收到的语义信息在接收端完成各种下游任务。在面向任务的图像语义通信中,常见的智能任务包括分类、检索和分割。例如,Kang 等人提出了一种基于语义三元组的无人机节能语义通信框架,其语义编码器根据终端用户需求进行个性化[10]。此外,Pan 等人提出了一种面向图像分割的车联网语义通信系统,其中重构的语义特征用于接收端图像分割[11]。

另一方面,面向图像重建的语义通信侧重于原始图像的传输和恢复。早期的语义通信研究多基于信源信道联合编码(Joint Source-Channel Coding,JSCC)方法。随后,DeepJSCC 作为一种基于深度神经网络的 JSCC 方法,在低信噪比(Signal-to-Noise Ratio,SNR)和低压缩比(Compression Ratios,CRs)场景中获得出色的性能[12]。与此同时,研究者基于 JSCC 方法开发了各种各样的语义通信模型。例如,Xu 等人通过引入基于注意力的 JSCC 方法,帮助模型适应不同 SNR 下的无线通信环境[13]。Yang 等人开发了一种可根据信道条件和图像内容动态调整特征的长度的 JSCC 模型,以此来节约带宽资源[14]。Wu 等人提出了一种信道去噪扩散模型,用于无线信道语义通信,通过利用扩散模型消除噪声的特性,提高了常规 JSCC 模型的图像还原能力[15]。

近期,一些学者尝试从通信时的编码和解码方式入手,改进图像语义通信模型的性能。例如,Hu 等人提出了一种基于掩码自编码器和矢量量化变分自编码器模型的语义通信系统,通过共享离散的码本显著降低通信开销[16]。在此基础上,Miao 等人提出一种语义通信系统 VQ-DeepSC-E,该系统利用生成对抗网络(Generative Adversarial Network,GAN)提取图像的特征,减少了重建图像时的伪影现象[17]。这种共享码本的方法极大地减少了带宽消耗,然而码本的索引在传输的过程中易受到噪声的干扰。

目前,多数面向图像重建的无监督语义通信系统在极低的 CRs 下会出现各类图像退化问题,如模糊、伪影或棋盘化等[18]。此外,由于高度压缩图像与人类视觉感知之间存在巨大差异,这些图像可能无法直接适用于图像分类、对象检测和语义分割等下游任务[19]。

为了缓解上述图像退化问题,部分专家学者提出了生成式语义通信(Generative Semantic Communication,GSC)方法。GSC 方法主要利用能够抽象概括图像信息的高层语义,如文本信息和语义分割信息,控制语义解码器生成符合要求的图像,并将原本无监督的训练方法转为有监督方法[20]。例如,Huang 等人提出了一种基于强化学习的语义通信模型,该模型使用语义分割图像作为语义信息,根据下游任务自适应地选择将要传输的语义信息[19]。Lokumarambage 等人将风格图像设置为共享的先验知识,在发送端依据该知识生成语义分割图,在接收端依据该知识和语义分割图生成结果图像[21]。Grassucci 等人在发送端提取并传输多张 One-hot 编码的语义分割图,在接收端去噪后这些图被用于引导扩散模型生成对应的图像[22]。Lee 等人提出了一种能够节约能源的通信框架,用文本替代了原本将传输的图像,并让用户在本地根据文本信息生成图像[23]。上述研究展示了高层语义在图像语义信息压缩和表达方面的优势。然而,大部分 GSC 方法依赖于人工标注数据或预训练的模型,无法从原始图像中自动提取高层语义信息。相比之下,对比学习作为一种自监督学习方法,可以获得适用于各种下游任务的通用特征,而不依赖于人工标记。此外,对比学习代理任务的目标是生成能够与其他图像区分的编码,即将每张图像单独视为一个类别,因此提取出的特征可以近似为这张图的专属标签。并且,Caron 等人的研究表明,对比学习训练时会关注与图像语义分割相关的信息[24]。因此,本文尝试将对比学习模型提取的特征视为一种高维标签,取代依赖人工标注的语义分割图、文本信息或分类信息,从而满足低压缩比情况下图像重建过程中对大量高层语义的需求。

鉴于高层语义信息较为抽象,为建立其到图像像素信息之间的映射,需要相关模型具备强大的生成能力来填补高层语义缺失的细节。目前,扩散模型在图像生成领域的研究日渐增加。相较于 GAN,扩散模型在图像生成质量和多样性

方面具有显著优势[25]。此外，它可以通过添加引导信息指导生成过程，以确保生成的图像满足特定需求[26]。

为解决上述问题，本文的主要工作如下：

(1)本文设计了一种基于对比学习的语义编码器(Momentum Contrast Semantic Encoder，MoCoSE)。MoCoSE 通过改进训练自监督学习过程，使模型能够更好地适应通信环境。使得编码器能够通过自监督学习从图像中提取高层语义特征。在训练中，结合变分自编码器(Variational Autoencoder，VAE)模型让编码器能够提取低层语义特征。与其他利用高层语义特征的模型相比，MoCoSE 不依赖人工标签进行训练。

(2)本文利用扩散模型建立高维标签到图像像素的映射。研究表明，在语义通信过程中，即使高层语义被噪声严重干扰，使用这些语义特征引导的扩散模型仍然可以生成具有与原图相同核心语义的图像。与传统的无监督语义通信方法相比，该模型能够有效缓解图像退化问题，从而提高重建图像的视觉质量。

(3)实验结果表明，在低 CRs 条件下，与传统通信和四种无监督图像语义通信模型相比，本文所提模型在传输图像核心语义信息方面优势显著，其生成的图像表现出更好的主观视觉质量。此外，MoCoSE 提取的高层语义信息相比低层语义信息具有更佳的抗噪声能力。

2 预备知识

本节分别介绍对比学习和扩散模型相关知识。

2.1 对比学习

在深度学习中数据量一直是制约模型性能的一个关键因素，因此如何利用大量且易于获得的无标注数据一直是深度学习研究中的重要问题[27]。对比学习是一种基于样本间差异的自监督学习方法。其由于不需要人工标签的特性，常用于大模型的预训练和未标记数据集的特征提取。

目前，对比学习中主要的代理任务是个体判别任务(Instance Discrimination，InstDisc)[28]。个体判别任务的提出者认为，让有监督模型能够区分相似图像与非相似图像的能力来源于数据本身，而不是语义标签。因此，个体识别任务在训练时会将每个样本当作一个单独的类别，并赋予一个伪标签，然后通过伪标签训练模型区分每个类别间的区别。在这个过程中会得到图像特征编码，并使得相似图像的特征在超球面上尽可能地接近，而不相似图像的特征在超球面上尽可能地远离。这种方式使得模型将每一张图视为单独的一个类别。使用正则损失互信息噪声对比估计(Info Noise Contrastive Estimation，InfoNCE)损失函数可以描述该任务的训练过程[29]。给定一组总数为 K 的样本编码 $\{k_0, k_1, \cdots, k_K\}$ 集合和查询样本编码为 q，设 k_+ 为 q 的正样本，则损失函数可以表示为

$$L_{\text{Info}} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)} \tag{1}$$

其中，τ 表示调控训练难度的温度超参数。

对比学习模型的训练通常需要大量的迭代次数和负样本，以获得理想效果。为降低训练过程中对显存和时间的需求，本文选择了 MoCo 模型作为语义编码器的基础模型[30]。MoCo 模型主要由编码器和动量编码器组成，其通过缓慢更新动量编码器权重，使动量编码器输出的特征与编码器输出的特征保持一致的同时，提供更多差异化信息以增强对比效果，从而在较小批次和更少迭代轮次下实现理想效果。

2.2 扩散模型

扩散模型是一种生成式模型，它的特点是通过逐步去除噪声来生成图像，而不是直接生成结果图像[28]。扩散模型的训练可以分为前向过程和逆向过程两个部分[31]。前向过程是一个将高斯噪声逐步加入数据中的过程，可以用公式表示为

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \boldsymbol{I}) \tag{2}$$

其中，x_t 表示过程中第 t 步的数据，β_t 表示提前设定的方差超参数，$\boldsymbol{I}$ 表示单位矩阵。当 t 足够大时可以认为 x_t 是服从高斯分布的噪声。当 x_0 表示原始数据，$\alpha_t = 1-\beta_t$ 且 $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ 时，公式可以表示为

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \tag{3}$$

逆向过程从 x_t 开始逐步去除噪声获得 x_0，这个过程可以使用神经网络进行拟合，表示为

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \epsilon_\theta(x_t, t), \sigma_t^2 \boldsymbol{I}) \tag{4}$$

当 $\sigma_t^2 = \dfrac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ 时，这个过程的损失函数可以表示为

$$L_{\text{sample}} = \mathbb{E}_{x_0, \epsilon}\left[\left\| \epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, t\right) \right\|^2\right] \tag{5}$$

其中，ϵ_θ 表示参数化神经网络计算出的噪声。

扩散模型的生成可以分为无条件生成和条件生成两种，其中条件生成又可以分为分类器引导生成[25]和无分类器引导[26]生成两种。分类器引导生成需要额外训练一个分类器，这会增加训练开销。无分类器引导生成，不需要额外的结构也可以进行条件生成，并且可以根据需求调整语义信息在生成图像中的权重，但需要同时训练模型的无条件生成和条件生成。损失函数可以表示为

$$L_{cfg}=\mathbb{E}_{x_0,\epsilon}[||\epsilon-\epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0+\sqrt{1-\bar{\alpha}_t}\,\epsilon,y_s,t)||^2] \tag{6}$$

其中，y_s 表示用于引导图像生成的语义信息。综合来看无分类器引导生成方式在效果和成本上优于分类器引导生成方式。

此外，当前的扩散模型通常采用 U-Net 模型作为主干进行图像生成。为提升模型的扩展性及其对多模态信息的适应能力，一些学者提出了多种基于 Vit(Vision Transformer)的扩散模型。本文选择其中的 U-Vit 模型作为语义解码器的基础模型[32]。U-Vit 模型基于 Transformer 模块，通过在模块间引入长距离跳跃连接来模拟 U-Net 的 U 型网络结构，其在扩散模型中表现出优于传统 U-Net 模型的性能。

3　模型设计与实现

本节主要简述所提模型的设计思路、结构与实现细节。

3.1　模型结构

一般情况下，人类在有目的拍摄或绘制除风景以外的图像时，会选取某些特定的事物作为目标进行构图并占据图像的主要位置。这些事物是人拍摄或绘制的目标，并且往往可以被分类。因此，本文将这类事物的分类结果作为一张图像的核心语义。本文主要考虑低 CRs 和低 SNR 通信条件下以核心语义作为主要目标的一对一图像语义通信场景。与通常考虑重建视觉效果一致的图像语义通信模型不同，本文所提模型的主要目标是在接收端重建与原始图像核心语义一致的图像，而不是简单地追求图像的像素一致。

本文模型运行时的结构如图 1 所示，主要由 3 个部分构成：发送端、物理信道和接收端。发送端由语义编码器和信道编码器组成。语义编码器负责提取语义特征，其中包含高层语义和低层语义。信道编码器则根据压缩比将原始语义特征进行信道编码。假设输入系统的原始图像为 $x\in R^{B\times C\times H\times W}$，其中 B 表示批大小，C 表示图像的通道数，H、W 分别表示高度和宽度。发送端的图像语义通信过程可以表示为

$$z_x=C_e(S_e(x)) \tag{7}$$

其中，z_x 表示功率归一化后将要发送的语义信息，$S_e(\cdot)$ 表示语义编码器，$C_e(\cdot)$ 表示信道编码器。发送信息的 CRs 的计算过程可以表示为

$$\mathrm{CRs}=\frac{N\times 32}{C\times H\times W\times 8} \tag{8}$$

其中，N 表示发送信息的维度。

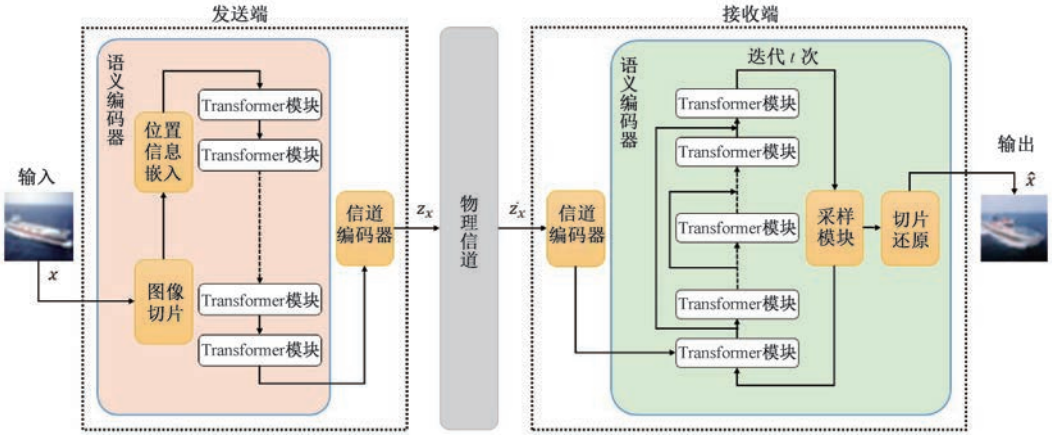

图 1　本文提出的自监督图像传输语义通信模型结构图

物理信道部分由通信设备构成。主要负责将 z_x 从发送端传输到接收端。本文主要考虑加性高斯白噪声(Additive White Gaussian Noise，AWGN)信道、瑞利(Rayleigh)信道和莱斯(Rician)信道，通信过程可以表示为

$$\bar{z}_x=hz_x+\mathrm{n} \tag{9}$$

其中，$\bar{z}_x$ 表示受到噪声干扰后的 z_x，h 表示信道增益，n 表示服从 $\mathcal{N}(0,\sigma^2)$ 的 AWGN。

接收端由信道解码器和语义解码器组成。信道解码器负责在最小化噪声的条件下，将由信道编码器编码的语义信息恢复到原始语义信息的大小。而语义解码器负责根据恢复的语义信息，生成具有语义一致性特征的图像。在接收端生成图像的过程可以表示为

$$\hat{x} = S_d(C_d(\bar{z}_x)) \tag{10}$$

其中，$\hat{x}$ 表示在接收端重建的图像，$S_d(\cdot)$ 表示语义解码器，$C_d(\cdot)$ 表示信道解码器。

3.2 语义编码器

在无监督学习的情况下，人工标注的高层语义信息无法使用。因此，需要寻找一种能在该情况下使语义编码器能够提取图像高层语义特征的代理任务，以满足语义提取的优化目标。经过筛选，本文认为用于区分不同样本的个体判别任务能够满足提取图像高层语义特征这一目标。这是由于个体判别任务在训练时会计算不同特征之间的距离，并尽可能地缩小正样本特征间的距离。这一过程中提取出的特征，反映的是图像在模型构造的高维空间中的位置，并且具有相似语义的特征会聚合在一起。此外，高维空间中位置的差异也可以反映除核心语义外的其他语义信息。因此，本文认为这种高维特征可以作为一种仅供机器理解的高层语义，而人类要直接理解这种语义信息则需要训练专门的模型对其进行解码。

个体判别任务在训练时需要知道什么是相似的样本，因此如何构造合适的正负样本是影响模型性能的关键因素。本文发现，当前对比学习中的个体判别任务不需要模型准确表示方向信息，只需区分正负样本，这导致模型在方向表示能力上较弱。

VAE 模型本质上是一种生成式模型，其核心在于通过引入潜在变量 Z 实现对数据的整体表征。该模型在训练过程中以重构输入数据为目标，其可表示为 $p(X) = \sum_Z p(X \mid Z)p(Z)$。基于这一特性，VAE 的编码器在训练过程中会着重学习数据集的概率分布特征，从而获得更具代表性的潜在变量。为解决对比模型训练出的特征在图像结构表达方面弱的问题，需要再次引入低层语义信息来加强语义特征对于图像结构的表达能力，以提高生成图像与原始图像的相似度。为了避免高层语义和低层语义互相竞争资源（也即 bit 位），本文尝试划分一部分特征维度，使用 VAE 模型的解码器进行针对性训练，让特定特征维度能够专注于捕获低层语义信息。这种方法在保留高层语义特征抽象表征能力的同时，有效捕获了低层细节信息，从而实现了多层次语义特征的有机融合与优势互补。

为使上述两种模型完成混合训练，MoCoSE 的结构主要包含基础语义编码器、动量语义编码器和 VAE 解码器，如图 2 所示。其中，基础语义编码器与动量语义编码器的结构完全一致，主要结构均由多层 Transformer 模块堆叠而成。在语义编码器中，为了避免高层语义和低层语义在生成时互相干扰，在图像特征序列输入 Transformer 模块前，会插入两个固定的令牌分别用于聚合高层语义和低层语义的信息。对应位置的输出在经过线性层映射后，分别作为高层语义和低层语义的代表，用 q_m 和 q_v 表示。完成训练后，MoCoSE 只保留基础语义编码器，作为整体模型的语义编码器使用。语义编码器输出的是 q_m 和 q_v 拼接而成的混合语义。VAE 解码器中的主要结构同样由 Transformer 模块堆叠构成，输入混合语义中低层语义部分，输出拟合的图像。

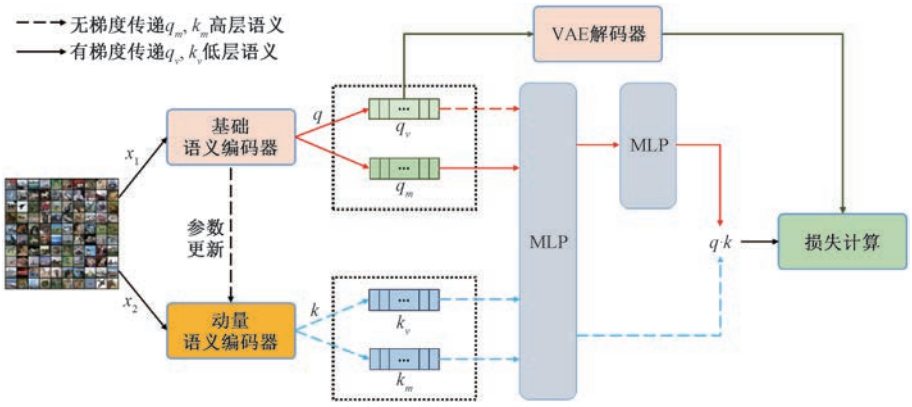

图 2 本文提出的 MoCoSE 训练过程示意图

在训练开始前，基础语义编码器和动量语义编码器中的参数将完全同步。随后，在每批次训练前，动量编码器先根据语义编码器的参数动量更新参数。在训练中，MoCoSE 对原始数据 x 进行两次随机数据增强，生成增强后的数据 $\{x_1, x_2\}$。其中，x_1 输入语义编码器中获得向量 q，而 x_2 则输入动量语义编码器中获得向量 k。将向量 q 截取固定长度得到 q_v，剩余部分为 q_m。将 q_v 输入 VAE 解码器以训练提取图像的细节信息，并令 q_v 只传递 VAE 解码器的梯度。q 经过两个多层感知机（Multilayer Perceptron，MLP）模块进行非线性变换，k 经过一个与语义编码器共享的 MLP 模块进行非线性变换。最终，映射后的 q 和 k 会同时输入 L_{Info}，用于优化模型。

一般而言，端到端语义通信模型直接使用受到通信噪声影响的语义信息进行训练。与有监督方法不同，对比学习不使用人工标注的标签，而是借助不同的 $\{q, k\}$ 对之间的差异来实现模型训练。然而，直接使用包含通信过程中过大噪声的特征会严重干扰模型对不同数据间差异的判断能力。本文通过比较原始语义信息与经过通信过程语义信息的差异，而不是比较受到噪声影响的 $\{q, k\}$ 对的方式，减少通信过程对模型的影响。定义损失函数为：

$$L_{SC} = L_{\text{Info}}(q, k) + L_{\text{Info}}(q_{SC}, k) + L_{\text{Info}}(q, k_{SC}) \quad (11)$$

其中，$\{q_{SC}, k_{SC}\}$ 表示信道解码器还原后的信息。

3.3　语义解码器

扩散模型是一种基于马尔科夫链的生成模型。与常见单步生成模型不同，通过多步生成获得高质量的图像。扩散模型不仅能够实现高精度的控制，还能生成丰富多样的结果。这种特性使得扩散模型在图像生成任务中具有广泛的应用潜力，尤其是在需要兼顾核心语义一致性和细节多样性的场景中。有效填补了高层语义信息引导模型生成图像时不足的细节信息。

为提升模型在无监督与低 CRs 情况下的性能，本文引入高层语义特征作为优化目标。由于高层语义信息具有高度压缩的特点，其常导致在人类视觉感知中不同的图像也提取出相似的高层语义信息。而一般端到端训练的语义解码器，在使用高度压缩到近似于分类标签的高层语义信息时难以生成符合语义且视觉质量足够的图像。因此，本文选择扩散模型作为语义解码器的基础模型。

扩散模型在训练时仅使用引导信息和扩散步数作为生成图像的依据。而在语义模型的训练过程中，通常会使用受到通信噪声干扰的语义信息。使用这些受污染的语义信息训练扩散模型可能阻碍模型的收敛，增加训练难度。针对这一问题，本文将 SNR 信息嵌入，使得扩散模型能够根据不同的 SNR 信息，对语义信息进行有选择地接受。该方法的损失函数表示为

$$L_D = \mathbb{E}_{x_0, \epsilon}[\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, y_{snr}, y_{ns}, t) \|^2] \quad (12)$$

其中，y_{snr} 表示通信时的 SNR 信息，y_{ns} 表示混入通信噪声的语义信息。

为使 U-Vit 模型适应上述需求，本文在其输入中添加了两类信息：语义信息和 SNR 信息。语义信息在输入阶段通过线性层映射到与 Transformer 模块匹配的维度。SNR 信息则在转换为噪声功率后，采用正余弦嵌入方式进行编码。

3.4　信道编解码器

在本文提出的通信模型中，信道编解码器旨在增强模型对通信信道噪声的适应能力，让语义编码器与语义解码器在训练时专注于特征提取与特征解码，减少由噪声造成的语义精度损失。

一般情况下，语义通信模型在特定的 SNR 环境下训练，会影响模型在其他 SNR 环境下的泛化性能[13]。本文提出一种基于 Transformer 的信道编解码器，根据 SNR 信息动态调整对输入信息的取舍，从而增强模型在不同 SNR 环境下的适应能力和泛化能力。信道编解码器由基础 Transformer 模块构成，在运行与训练时，将语义编码与 SNR 信息作为输入。其中，SNR 信息通过正余弦嵌入的方式生成特定的特征。

3.5　两阶段训练过程

本文选择基于对比学习的模型作为语义编码器，通过自监督学习提取高层语义。这种语义编码器与 JSCC 方法中的语义编码器不同，可以在不使用语义解码器的情况下单独完成训练。另外，为将高层语义解码为像素，本文提出基于扩散模型的语义解码器。然而，扩散模型进行无分类器引导训练时，需要较为稳定的输入信息作为引导，因此无法直接与编码器进行联合训练。

为有机结合对比学习模型与扩散模型，并显著降低模型训练对显存的需求，本文尝试将传统的端到端训练过程，分为两个独立的阶段，即“语义提取”和“语义解码”。如图 3(a)所示，语义提取阶段的主

要目标是对发送端的语义编码器、信道编码器以及接收端的信道解码器进行训练，旨在提升模型对语义进行联合编码的能力。在此过程中，仅信道编解码器能够获取 SNR 信息。该阶段主要用 L_{SC} 来计算模型的损失。如图 3(b)所示，语义解码阶段的主要目标是对接收端的语义解码器进行训练，发送端的语义编码器和信道编码器以及接收端的信道解码器的参数被冻结。此过程用从上一阶段训练出的语义特征，训练接收端的语义解码器，用损失函数 L_D 计算模型的损失。

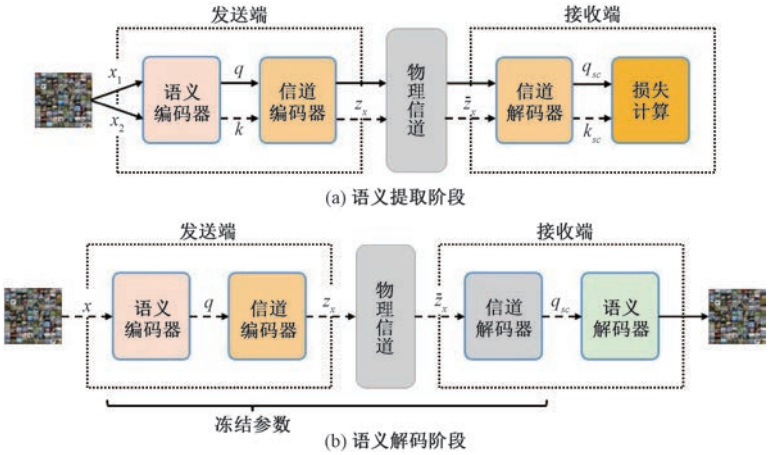

图 3 本文提出的模型不同训练阶段示意图

4 实 验

在本节中，将本文提出的模型与其他通信方法进行比较，并对各个模块进行测试。

4.1 仿真设置

在本小节将对仿真实验进行设定。

4.1.1 训练设置

本文使用 CIFAR10 数据集与 Tiny-Imagenet 数据集进行模型的训练与测试。CIFAR10 数据集包含 60 000 张 32×32 像素的彩色图像，涵盖 10 个类别，分为训练集和测试集两个部分。Tiny-ImageNet 数据集包含 120 000 张 64×64 像素的彩色图像，涵盖 200 个类别，分为训练集、测试集和验证集三个部分。

本文模型发送端的语义编码器由 12 层 384 维的 Transformer 模块构成。其注意力模块包含 12 个注意力头，使用偏置项调整输出的偏移量，设置 MLP 模块中隐藏层维度为输入维度的 4 倍。在训练 CIFAR10 数据集时，图像被切分为 4×4 大小的补丁输入模型，输出的语义信息通过 MLP 模块转换为 128 维。在训练 Tiny-ImageNet 数据集时，加载基于 ImageNet 预训练的对比学习模型，将图像放大为 224×224 并切分为 16×16 的补丁输入模型，输出的语义信息通过 MLP 模块转换为 256 维。负责对语义编码器进行低层语义训练的 VAE 解码器由 8 层 384 维的 Transformer 模块组成。其注意力模块包含 12 个注意力头，使用偏置项调整输出的偏移量，设置 MLP 模块中隐藏层维度为输入维度的 4 倍，采用 U-Vit 模型中的长跳跃结构。

信道编解码器由 5 层 384 维的 Transformer 模块构成。其注意力模块包含 6 个注意力头，使用偏置项调整输出的偏移量，设置 MLP 模块的中间维度扩展比例为 4，采用 U-Vit 模型中的长跳跃结构。

接收端的语义解码器基于 U-Vit 模型构建，主要由多层 Transformer 模块组成。在 CIFAR10 数据集上训练语义解码能力时，将图像切分为 4×4 大小的补丁输入模型。此时 Transformer 模块为 12 层 512 维，注意力模块包含 8 个头，禁用偏置项调整注意力模块的输出，设置 MLP 模块中隐藏层维度为输入维度的 4 倍。在 Tiny-ImageNet 数据集上训练语义解码能力时，将图像切分为 4×4 大小的补丁输入模型。此时 Transformer 模块为 16 层 768 维，注意力模块包含 12 个头，禁用偏置项调整注意力模块的输出，设置 MLP 模块中隐藏层维度为输入维度的 4 倍。

本文选用 AdamW(Adaptive Moment Estimation Weight Decay)优化器，令语义编码训练时的学习率为 0.0001，动量 $\beta_1=0.9$，$\beta_2=0.95$；令语义解码器训练时的学习率为 0.0001，动量 $\beta 1=0.99$，$\beta_2=0.999$。令语义编码器的权重滑动平均更新率为 0.99，语义解码器的权重滑动平均更新率

为 0.9999。

为优化模型的训练效率和性能，本文采用分阶段训练策略。在“语义提取”阶段，先使用传统对比学习方法对语义编码器进行独立预训练，使其充分学习高层语义特征。随后，结合信道编解码器，在 SNR 为－10dB 到 20dB 的 AWGN 信道环境下，利用本文提出的方法进行联合训练。这一较宽的 SNR 范围旨在使语义编码器和信道编解码器能够适应多样化的信道条件，从而增强其鲁棒性。在“语义解码”阶段，固定语义编码器和信道编解码器的参数，并在 SNR 为 0dB 到 20dB 的 AWGN 信道环境下训练语义解码器。由于语义解码器是基于扩散模型构建的，其完全依赖于接收到的信息进行生成引导，并需要根据关键信息补充缺失内容，因此对噪声更为敏感。相对较高的 SNR 范围能够为其提供更稳定的训练环境，确保其生成质量。随后，在 Rayleigh 和 Rician 信道环境下进行迁移训练。

为了详细分析模型在不同语义情况下的表现，本文在语义编码器结构和参数量相同的情况下，测试了模型在多种情况下的性能。

(1)Mix-Diff：使用 MoCoSE 作为语义编码器提取混合语义，并使用对应训练的扩散模型进行图像的生成。

(2)VAE-Diff：使用 VAE 模型作为语义编码器提取低层语义，并使用对应训练的扩散模型进行图像的生成。具体而言，在 MoCoSE 的基础上，移除了对比学习相关内容，仅保留与 VAE 相关的训练过程。

(3)MoCo-Diff：使用 MoCoSE 作为语义编码器提取高层语义，并使用对应训练的扩散模型进行图像的生成。具体而言，在 MoCoSE 的基础上，移除了 VAE 相关内容，仅保留与对比学习相关的训练过程。

(4)Lable-Diff：使用标签信息作为高层语义直接输入语义解码器，替代模型中语义编码器和语义通信部分。用于评价模型在语义传递和语义提取方面的综合性能。

4.1.2　对照模型

为了全面评估本文模型的性能，选取了一种传统通信方法、四种无监督图像语义通信模型作为对比方法，如下：

(1)BPG＋LDPC：图像采用 BPG(Better Portable Graphics)格式进行压缩编码。在通信时采用 802.11－2020 标准中的 LDPC(Low Density Parity Check)编码方案进行信道编码，编码长度固定为 1944，码率设置为 1/2。在传输过程中使用 QAM (Quadrature Amplitude Modulation)对信道编码进行调制解调，符号表大小设置为 16。此外，传输失败的图像由相同尺寸的黑色图像代替。

(2) DeepJSCC[12]：DeepJSCC 是较早使用了 JSCC 方法的图像语义通信模型，用卷积神经网络(Convolutional Neural Network，CNN)模型提取与解析图像的语义信息。

(3)DeepJSCC-V[33]：DeepJSCC-V 是基于 CNN 支持可变码长的 JSCC 模型，通过额外的策略网络来控制发送码长。

(4) WITT(Wireless Image Transmission Transformer)[34]：WITT 模型采用 Swin Transformers 模块提高语义通信模型获取全局信息的效率，克服了传统 CNN 模型在处理高分辨率图像时性能显著下降的局限性。

(5)CDDM(Channel Denoising Diffusion Models)[15]：CDDM 利用扩散模型在去噪方面的优异性能，对接收到的编码信息进行去噪还原，并用经过专门对 CDDM 进行适配训练的语义解码器进行解码。

以上无监督图像语义通信模型，首先在 SNR 为－10dB 到 20dB 的 AWGN 信道环境中进行训练，然后在 Rayleigh 和 Rician 信道上进行迁移训练。并且为了体现这些模型在低 CRs 情况下的性能，在通信时的维度与本文模型保持一致。

4.1.3　评估指标

目前常用的图像评价指标主要依赖像素间的相似度进行评价，例如 PSNR(Peak Signal-to-Noise Ratio)和 SSIM(Structural Similarity)。然而，这类指标并不能衡量语义层面的变化。为此，本文从语义还原、像素还原和感知质量三个角度综合量化重建模型的性能。在语义还原方面，选用通过在对应数据集上微调的在 ImageNet 数据集上预训练 Vit 模型的图像分类正确率作为指标。若分类模型能正确识别被测模型生成的图像，则判定被测模型成功传递原始图像的核心语义信息。在像素还原方面，保留传统指标 SSIM，衡量生成图像与原始图像在像素结构上的相似性。在感知质量方面，选用 FID (Fréchet Inception Distance)分数，通过计算图像特征间的距离，可以量化生成图像的感知质量，衡量生成图像的感知质量是否与原图相近[35]。

4.2　语义表现

本文采用了微调后的 Vit 模型来衡量生成图像

的核心语义。图 4 展示了不同通信条件下，使用 Vit 分类模型对生成图像进行分类时的正确率比较。可以看出传统通信方式在 SNR 相对较高时表现最好，但随着 SNR 的下降，传统通信方式的性能出现了断崖式下跌。相比之下，本文提出的模型在 SNR 较低的条件下展现出显著优势，其分类正确率相对于传统方式和基于低层语义的通信模型而言有显著提升，能够更精准地传递图像核心语义，并生成易于识别的高质量图像。这就充分体现在低 CRs 的场景下，本文模型在语义传递与还原能力上的优越性，也证明了混合语义相对于低层语义，具有了更好的抗噪能力和语义表达能力。

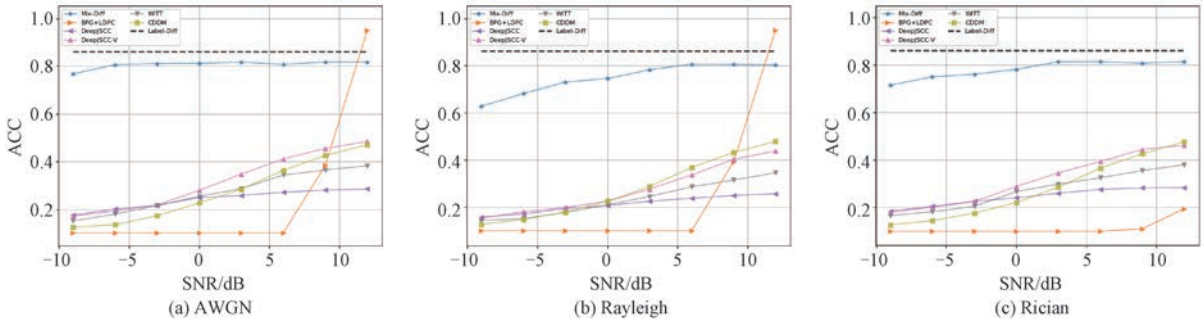

图 4 不同方法下分类正确率随 SNR 变化的趋势

4.3 图像质量

由于通过第三方分类模型进行分类，只能判断图像中的核心语义信息是否能够被正确识别，无法评估模型生成图像的相似度和质量。因此，本节使用 FID 分数和主观视觉质量来评价生成图像与原图特征的相似度以及视觉质量。

4.3.1 客观质量

为了评估生成图像的质量，本文通过 SSIM 评分衡量生成图像与原始图像在像素级的相似度，通过 FID 分数衡量原始图像与生成图像的特征在高维空间中的距离。如图 5 所示，当 SNR 较高时，传统通信方式具有与原图基本一致的视觉表现力，但随着 SNR 下降，图像质量快速下降，直到全部图像都无法辨认。此外，Mix-Diff 和 MoCo-Diff 在两种图像质量指标上都优于 Label-Diff。这表明本文模型提取的高层语义，不仅包含标签信息，还包含额外的信息。值得注意的是，尽管本文模型在低 CRs 时生成图像的视觉质量显著优于对比模型，但其 SSIM 值并未同步提升。下面分析原因：首先，高层语义信息是图像信息的高度概括，在下游任务无需过多细节时，会自然忽略部分冗余信息；其次，由于扩散模型的生成具有多样性，在核心语义一致的情况下使得与原始图像间存在不同的细节，导致生成图像的像素相似度降低。

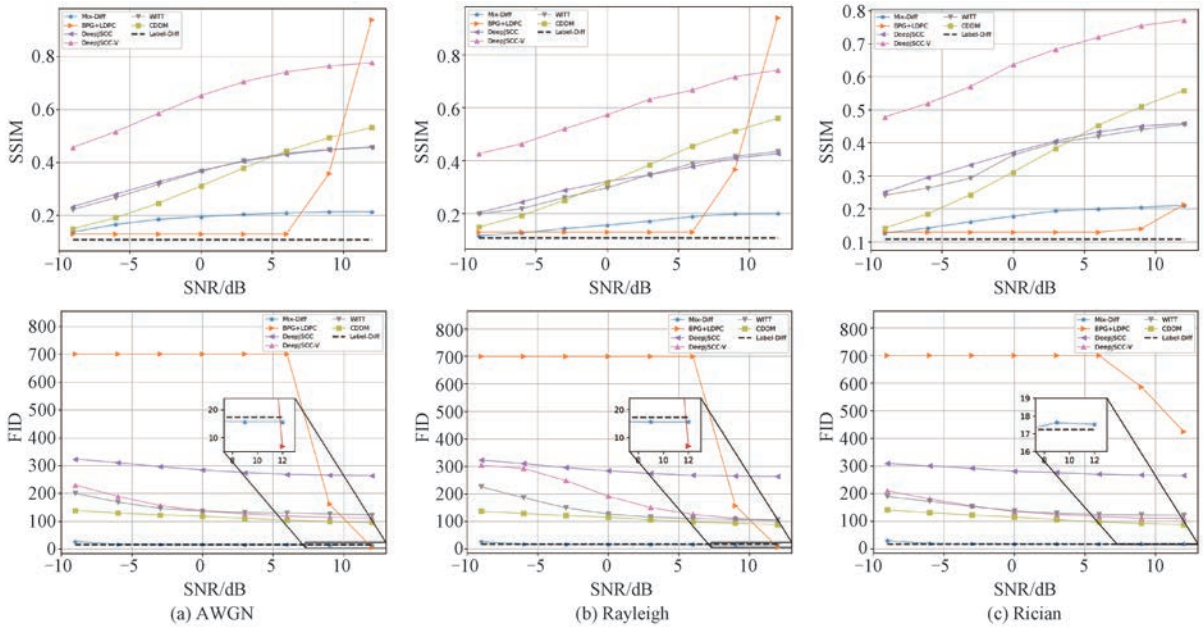

图 5 不同方法下像素相似度和视觉质量随 SNR 变化的趋势

4.3.2　主观质量

图 6 中提供了 CIFAR10 数据集上部分图像的原图，以及在不同传输方法和不同 SNR 下生成图像的可视化表示。可以看出，在 SNR 较低时传统通信方法传输的图像数据大多不可辨识。而本文所提出的模型在低 CRs 条件下具有优越的生成性能，可以轻易地判断图像中的核心语义信息。而对照模型在此条件下生成的图像质量较差，难以识别其中可能包含的语义信息。

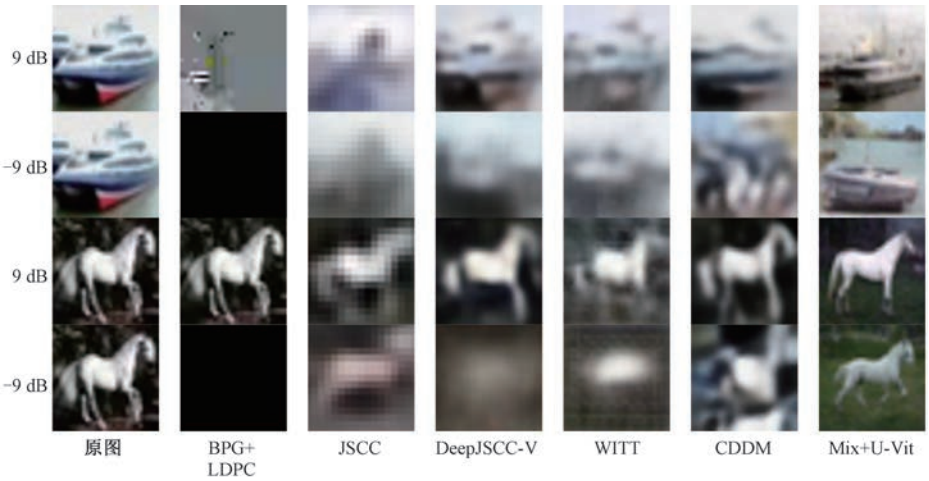

图 6　AWGN 信道下不同方法不同 SNR 情况下的视觉效果对比

此外，结合图 4 和图 6 的内容可以看出，传统模型在低 CRs 条件下为了追求训练指标的最大化，会牺牲生成图像的细节信息，通常表现为在图像中生成模糊的色块。这种方式极大地降低了生成图像的视觉质量。此外，由于这类图像缺乏细节和信息量，容易被人或模型错误识别为其他事物。因此，本文认为在低 CRs 和低 SNR 的极端情况下，高层语义信息相比于低层语义能更好地传递核心语义。这种情况下，应该提取图像的高层语义信息，然后结合具有少量细节的低层语义信息，以生成的方式重构图像，从而最大限度地提高生成图像的核心语义的准确度和视觉表现。

4.4　性能分析

在这个部分将对影响模型性能的高低层语义比例和生成时采样步数进行分析。

4.4.1　语义比例

为了探究高层语义和低层语义不同维度的比例对模型性能的影响，调整了对比学习训练和 VAE 模型训练时的维度的比例。本文主要对三个参数比值进行测试，将 128 个维度分为 32、64 和 96 个维度进行 VAE 模型训练。本文还使用对比学习模型和 VAE 模型作为语义编码器，分别表示纯高层语义和纯低层语义，以便更好地分析不同层次语义信息的差异。在训练混合语义时，VAE 模型主要训练提取物体结构和位置等低层语义信息的能力，其余部分由对比学习进行提取，因此使用黑白图像作为优化目标。而纯 VAE 模型需要独自提取并传递其能获取的所有语义信息，因此使用彩色图像作为优化目标。

如图 7 所示，高层语义具有领先于低层语义的核心语义传递与表达能力，而在像素相似度上低层语义相对于高层语义信息有着明显的优势。在 SNR 较高时，高层语义和低层语义生成图像的质量相差不大。随着 SNR 的下降，低层语义在图像质量与像素相似度上的性能显著下降，而高层语义则保持相对平稳的下降趋势。这证明了基于对比学习提取出的高维标签相对于低层语义具有更好的抗噪性。

此外，混合语义的表现证明，在信息量足够的情况下，可以将高层语义信息和低层语义信息混合，并在不影响图像核心语义信息表达的情况下，提高生成图像与原始图像的相似度。此外，生成图像与原始图像之间的像素相似度，也随着低层语义信息比例的增加而增加；在这种情况下，高层语义和低层语义比例相同的混合语义信息综合性能最好。但是，在带宽不足的情况下，混合大量的低层语义会影响高层语义的核心语义表达能力。

4.4.2　分辨率与压缩比

为探讨本文模型中分辨率与 CRs 对模型性能的影响，本文在更大规模的 Tiny-ImageNet 数据集上进行了测试。表 1 展示了本文模型与对照模型在两个数据集上的不同 CRs。其中，BPG＋LDPC 的 CRs 通过计算测试集图像传输时的平均 CRs 得出。测试结果如图 8 所示，结合图 4 和图 5 可以看出，

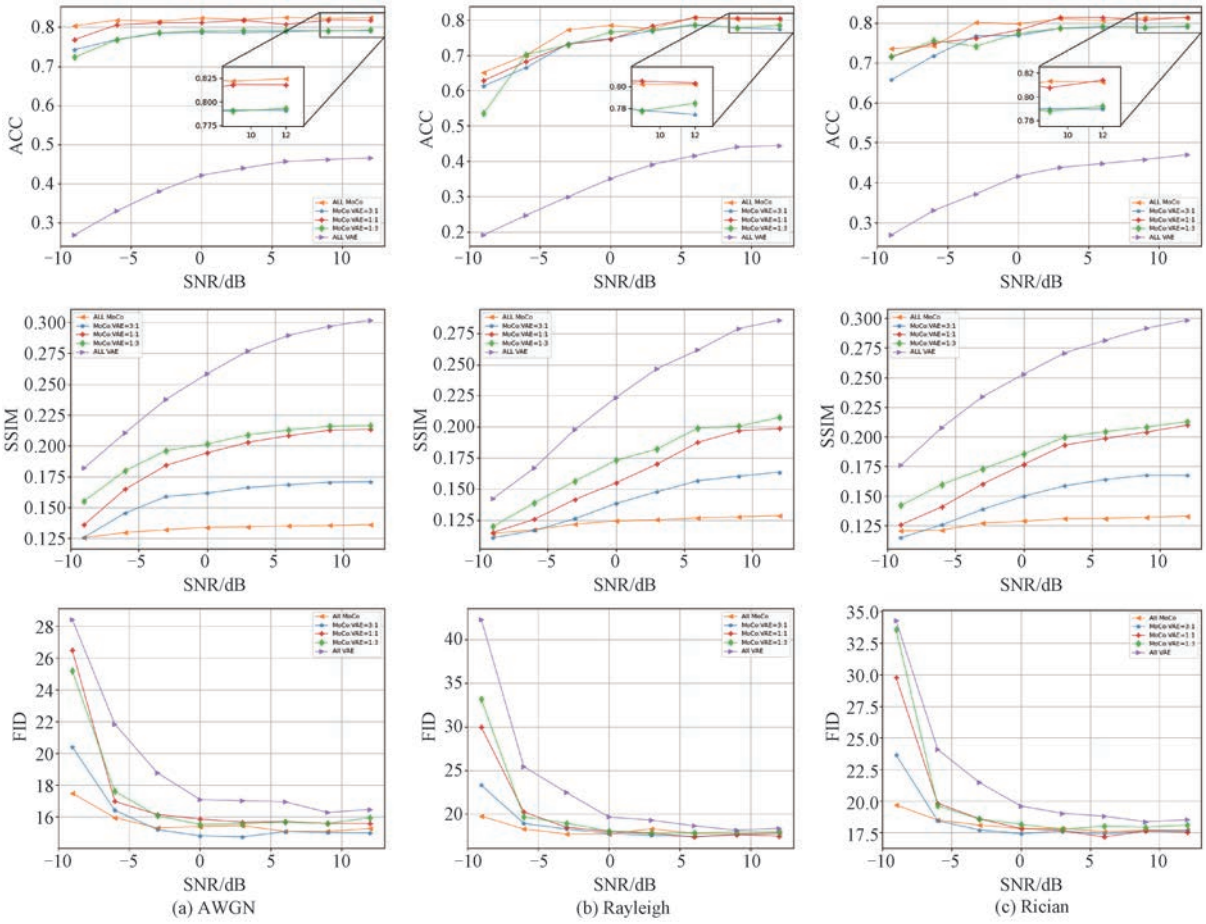

图 7 本文提出的模型在不同语义比例下性能指标随 SNR 变化的趋势

表 1 不同方法在两个数据集上的压缩率

方法	CIFAR10	Tiny-ImageNet
Mix-Diff	0.1667	0.0833
BPG+LDPC	**0.2247**	**0.2022**
DeepJSCC	0.1628	0.0833
DeepJSCC-V	0.1667	0.0833
WITT	0.1667	0.0833
CDDM	0.1667	0.0833

本文模型和对照模型在核心语义还原和像素相似度方面均表现出显著下降。基于 JSCC 的对照模型性能下降，主要是由于 CRs 降低导致低层语义信息的不足。MoCo-Diff、Mix-Diff 和 Label-Diff 的性能下降可能由于 CIFAR10 的训练集中每一类有 5000 个样本，而 Tiny-ImageNet 的训练集中每一类只有 500 个样本，每个类别的图像远少于 CIFAR10，且不同类别间的数据分布差异较大。这显著增加了语义解码器在使用高层语义进行训练时的拟合难度，导致其表现相比在 CIFAR10 上明显下降。此外，MoCo-Diff 在 SNR 为 12 时表现出强于 Label-Diff 的核心语义还原能力，也证明了当前的性能瓶颈主要源于语义解码器和数据集，而非对比学习的语义提取能力或相比在 CIFAR10 数据集上运行时更小的 CRs。因此，本文推测，解决语义解码器带来的性能

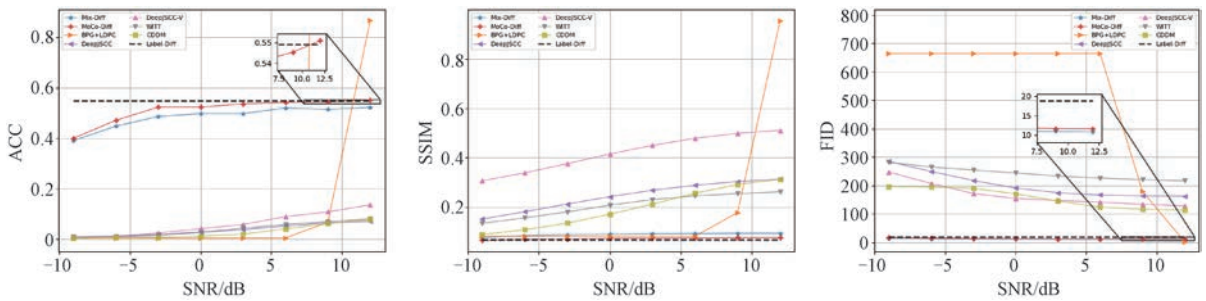

图 8 在 AWGN 信道与 Tiny-ImageNet 数据集上不同性能指标随 SNR 变化的趋势

瓶颈后,可以使用更小的 CRs 传输更大的图像,或在 CRs 不变的情况下,传输更多的低层语义信息以提高像素相似度。

4.4.3　采样步数

扩散模型在生成图像时需要进行逆向扩散过程,这个过程会根据采样步数的不同而消耗不同的时间。逆向扩散过程中采样步数的多少,会显著影响生成图像的质量、相似度以及生成所需的时间。如图 9 所示,当采样步数增加时,生成图像的核心语义表达、生成图像的质量以及生成图像与原图的相似度都随之提高。当采样步数为 50 步时,像素和特征的相似度比用 30 步和 10 步时有明显增加。然而,当采样步数增加到 100 步时,与 50 步相比,图像的质量和语义信息表达只有轻微提高,但时间消耗接近翻倍。如表 2 所示,在采样步数增加到 100 步时,时间消耗明显增加,但性能提升幅度有限,导致性价比较低。相比之下,采样步数为 50 步时性能较为均衡,综合性能最佳。

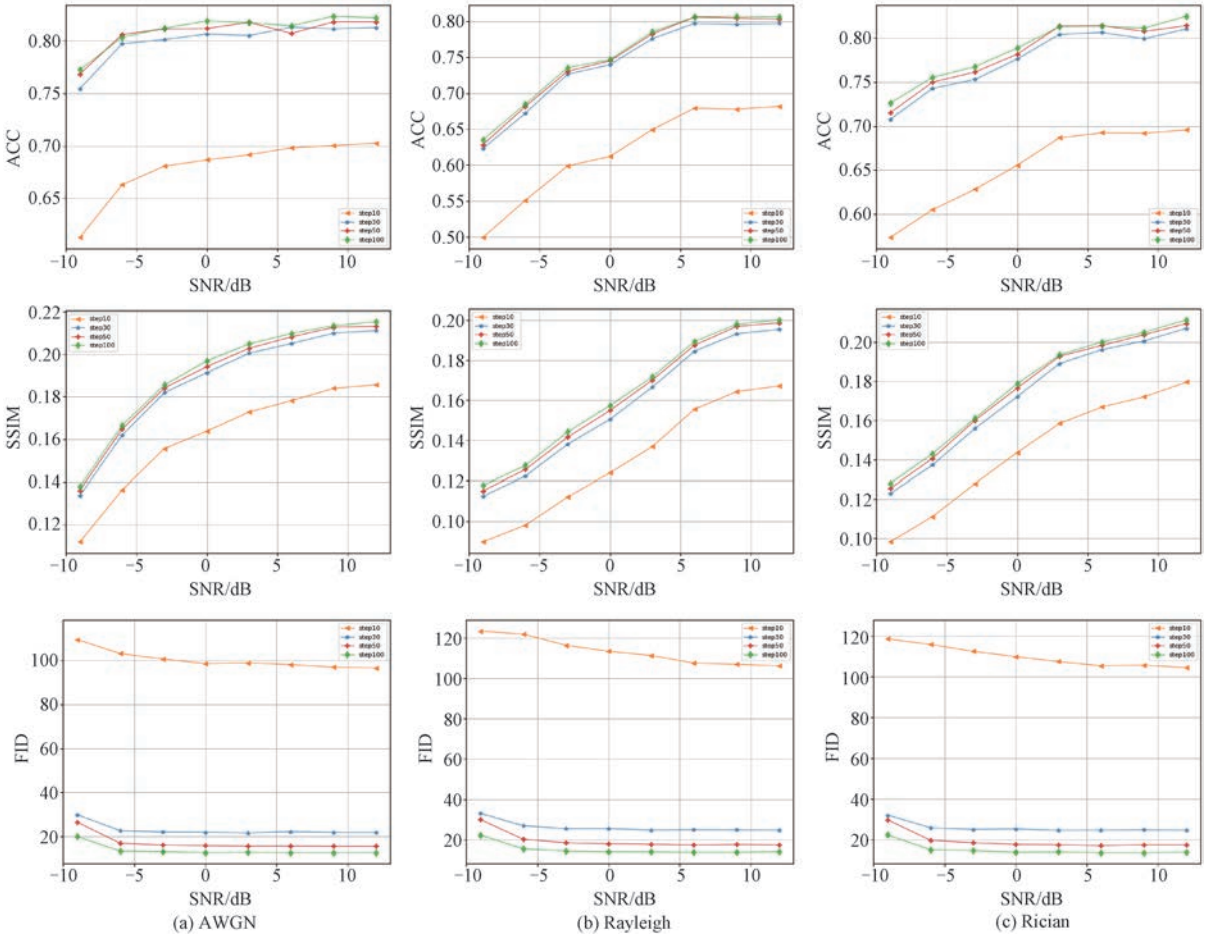

图 9　本文提出的模型在不同采样步数下性能指标随 SNR 变化的趋势

表 2　批次大小为 100 时不同采样步数平均时间

采样步数	平均时间/s
10	3.3523
30	10.2343
50	17.0935
100	33.9787

4.5　消融实验

为了验证本文所提出的模块和模型的有效性,将基于低层语义信息模型的四种语义编码器加入 U-Vit 作为语义解码器,进行联合训练。从图 10 的结果可以看出,采用低层语义的模型结合扩散模型后,可以通过牺牲部分像素的相似度,来提高生成图像核心语义的表达能力和质量。从图 10 中不同信道的表现可以看出,当底层语义信息充足时,使用扩散模型也可以很好地表示图像的核心语义。然而,与高层语义相比,低层语义信息的信息压缩程度较低,这导致在面对更复杂的语义时性能下降更为明显。通过对比不同 SNR 环境下的性能可以看出,高层语义具有更好的抗噪声性能和语义表达能力。这是由于低层语义通过描述大量的局部图像细节,间接表示图像的整体语义信息;一旦受到噪声的干扰,

就会导致内容的丢失或偏差，从而难以准确地推断图像的核心语义。另外，本文提取的高层语义信息包含全局和局部抽象语义，这使得部分信息的缺失不容易影响整体图像语义的表达。

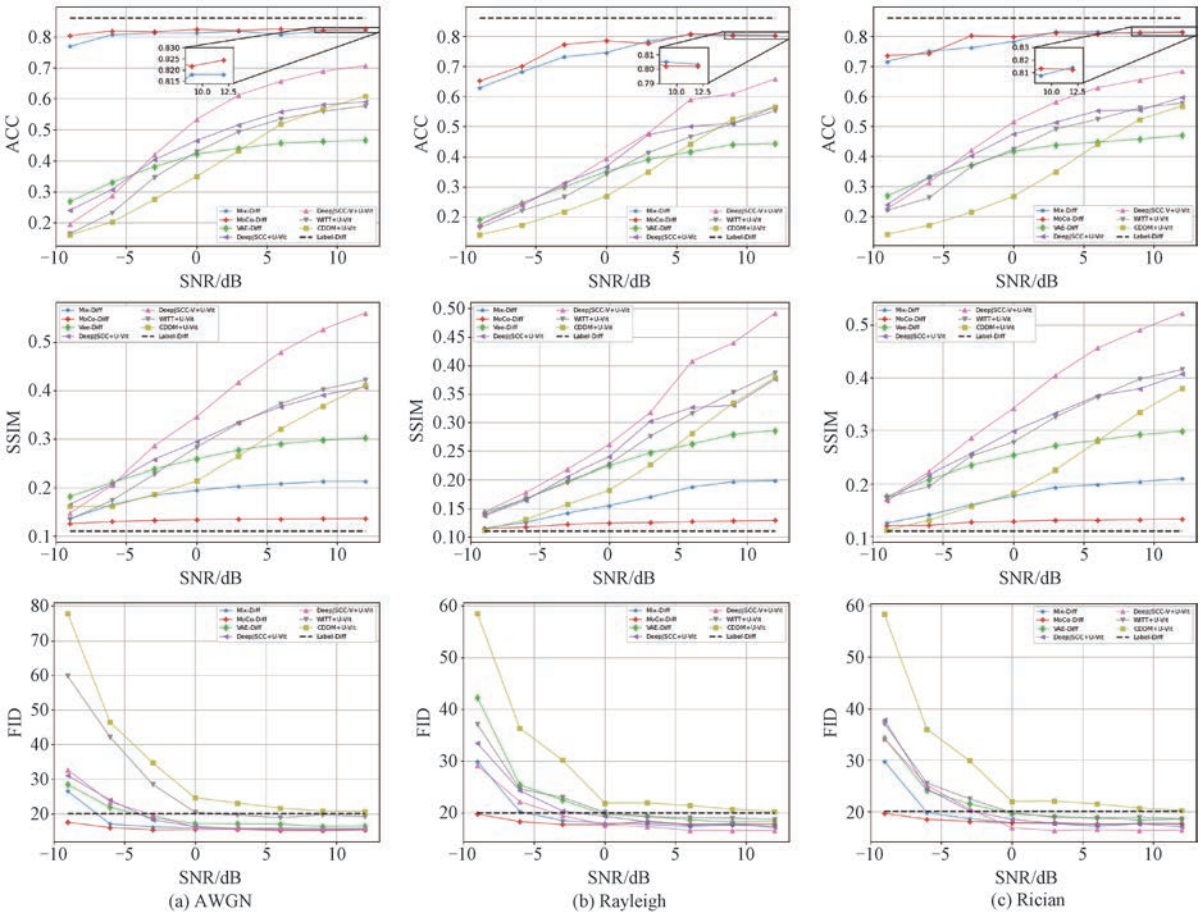

图 10 本文提出的模型与结合扩散模型后的对比模型在不同信道上性能指标随 SNR 变化的趋势

如图 11 所示，当基于低层语义模型的语义编码器使用本文的语义解码器进行图像生成，生成图像的视觉效果有了明显的改善。此外，四种使用扩散模型作为语义解码器的 JSCC 模型生成图像在视觉效果和语义表达能力上，随着 SNR 的降低而出现明显下降。这一现象验证了高层语义对于噪声具有更好的抵抗能力。这种抗噪性能的差异来源于少量的低层语义被噪声影响后，易被理解为其他高层语义。

图 11 AWGN 信道下本文提出的模型与结合扩散模型后的对比模型在不同 SNR 情况下的视觉效果对比

此外，为了验证 VAE 模型提取出的低层语义对于方位信息的提高，本文基于预训练的对比学习网络，训练了一个朝向判别器，来判断原始图像和目标图像的朝向是否一致。如表 3 所示，仅使用对比学习特征生成的图像基本损失了原图中所有方向信息。而在加入基于 VAE 的低层语义信息后，生成图像与原图的方向一致率有了显著提高。

表 3　AWGN 信道上 SNR 为 9 时生成图像与原图的朝向一致率

方法	朝向一致率(%)
水平翻转图像	1.29
MoCo-Diff	49.25
Mix-Diff	86.29
VAE-Diff	94.54

5　结　论

本文针对现有无监督语义通信模型在低 CRs 与低 SNR 条件下图像重建质量下降、有监督模型依赖人工标注数据的问题，提出一种能够利用高层语义的自监督语义编码器 MoCoSE，采用扩散模型提高生成图像的视觉质量与语义辨识度。仿真结果表明，在低 CRs 条件下，相较于对照的四种无监督语义通信模型，本文提出的模型具有更好的核心语义还原能力和图像生成质量。尤其是在低 SNR 条件下，本文提出的模型仍能正常工作，相对于传统通信方法和一般无监督语义通信模型具备更好的鲁棒性。此外，语义比例实验和消融实验的结果表明，高层语义比低层语义具备更好的抗噪能力与核心语义表达能力。

参　考　文　献

[1] Wheeler D and Natarajan B. Engineering semantic communication: A Survey. IEEE Access, 2023, 11: 13965-13995

[2] Shi G, Gao D, Song X, et al. A new communication paradigm: from bit accuracy to semantic fidelity. arXiv preprint arXiv:2101.12649, 2021

[3] Xie H, Qin Z, Li G Y, et al. Deep learning enabled semantic communication systems. IEEE Transactions on Signal Processing, 2021, 69: 2663-2675

[4] Han T, Yang Q, Shi Z, et al. Semantic-preserved communication system for highly efficient speech transmission. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 245-259

[5] Kang X, Song B, Guo J, et al. Task-oriented image transmission for scene classification in unmanned aerial systems. IEEE Transactions on Communications, 2022, 70(8): 5181-5192

[6] Jiang P, Wen C.-K, Jin S, et al. Wireless semantic communications for video conferencing. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 230-244

[7] Sheng Y, Li F, Liang L, et al. A multi-task semantic communication system for natural language processing//Proceedings of the IEEE 96th Vehicular Technology Conference (VTC2022-Fall), London, UK, 2022: 1-5

[8] Tang B, Li Q, Huang L, et al. Text semantic communication systems with sentence-level semantic fidelity//Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, UK, 2023: 1-6

[9] Qin Z, Tao X, Lu J, et al. Semantic communications: Principles and challenges. arXiv preprint arXiv:2201.01389, 2022

[10] Kang J, Du H, Li Z, et al. Personalized saliency in task-oriented semantic communications: image transmission and performance analysis. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 186-201

[11] Pan Q, Tong H, Lv J, et al. Image segmentation semantic communication over internet of vehicles//Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), Glasgow, UK, 2023: 1-6

[12] Bourtsoulatze E, Kurka D B, and Gündüz D. Deep joint source-channel coding for wireless image transmission//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019: 4774-4778

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

[14] Yang M and Kim H-S. Deep joint source-channel coding for wireless image transmission with adaptive rate control//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022: 5193-5197

[15] Wu T, Chen Z, He D, et al. CDDM: Channel denoising diffusion models for wireless semantic communications. IEEE Transactions on Wireless Communications, 2024, 23(9): 11168-11183

[16] Hu Q, Zhang G, Qin Z, et al. Robust semantic communications against semantic noise//Proceedings of the IEEE 96th Vehicular Technology Conference (VTC2022-Fall), London, UK, 2022: 1-6

[17] Miao Y, Yan J, Wang Y, et al. A semantic communication system based on vector quantization and generative model//Proceedings of the 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE), Guangzhou, China, 2024: 46-50

[18] Xing H, Kuang T, Feng L, et al. Image semantic communication framework with multi-target segmentation optimiza-

tion. chinese patent appl. CN117495998A, Feb. 2, 2024. (in Chinese)
(邢焕来,况田泽宇,冯力等. 一种多目标分段优化的图像语义通信框架, 中国, CN117495998A, 2024. 2. 2.)

[19] Huang D, Gao F, Tao X, et al. Toward semantic communications: deep learning-based image semantic coding. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 55-71

[20] Grassucci E, Mitsufuji Y, Zhang P, et al. Enhancing semantic communication with deep generative models: an overview//Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 2024: 13021-13025

[21] Lokumarambage M U, Gowrisetty V S S, Rezaei H, et al. Wireless end-to-end image transmission system using semantic communications. IEEE Access, 2023, 11: 37149-37163

[22] Grassucci E, Barbarossa S, Comminiello D. Generative semantic communication: diffusion models beyond bit recovery. arXiv preprint arXiv:2306.04321, 2023

[23] Lee H, Park J, Kim S, et al. Energy-efficient downlink semantic generative communication with text-to-image generators. arXiv preprint arXiv:2306.05041, 2023

[24] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021: 9630-9640

[25] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis//Proceedings of the Neural Information Processing Systems (NeurIPS), Virtual, 2021: 1-15

[26] Ho J and Salimans T. Classifier-free diffusion guidance//Proceedings of the Neural Information Processing Systems (NeurIPS), Virtual, 2021: 1-8

[27] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 9726-9735

[28] Wu Z, Xiong Y, Yu S X, et al. Unsupervised feature learning via non-parametric instance discrimination//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 3733-3742

[29] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018

[30] Chen X, Xie S, and He K. An empirical study of training self-supervised vision transformers//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021: 9620-9629

[31] Ho J, Jain A, and Abbeel P. Denoising diffusion probabilistic models//Proceedings of the Neural Information Processing Systems (NeurIPS), Virtual, 2020: 1-23

[32] Bao F, Nie S, Xue K, et al. All are Worth Words: A vit backbone for diffusion models//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023: 22669-22679

[33] Zhang W, Zhang H, Ma H, et al. Predictive and adaptive deep coding for wireless image transmission in semantic communication. IEEE Transactions on Wireless Communications, 2023, 22(8): 5486-5501

[34] Yang K, Wang S, Dai J, et al. WITT: A wireless image transmission transformer for semantic communications//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023: 1-5

[35] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium//Proceedings of the Neural Information Processing Systems (NeurIPS), Long Beach, USA, 2017: 1-12

XING Huan-Lai, Ph. D., associate professor, Ph. D. supervisor. His research interests include semantic communication, network function virtualization, software defined networks, artificial intelligence and evolutionary computation.

KUANG Tian-Ze-Yu, master student. His research interests include semantic communication and artificial intelligence.

XU Le-Xi, Ph. D., professor-level senior engineer. His research interests include big data, network analysis and intelligent operation.

LI Yang, Ph. D., assistant professor. Her research interests include semantic communication and edge intelligence.

ZHENG Dan-Yang, Ph. D., associate research professor and master supervisor. His research interests include network function virtualization, in-network computing, network reliability and safety, and digital twin systems.

LUO Shou-Xi, Ph. D., associate professor and master supervisor. His research interests include data center networks and networked systems.

DAI Peng-Lin, Ph. D., associate professor and Ph. D. supervisor. His research interests include intelligent transportation systems and vehicular cyber-physical systems.

LI Ke, Ph. D., lecturer and master supervisor. Her research interests lie in internet of vehicles.

FENG Li, Ph. D., Research professor and Ph. D. supervisor. His research interests include cyberspace security and artificial intelligence.

Background

Improving image transmission under constrained conditions is a current focal point in the image semantic communication research. Within this realm, the predominant approach is Joint Source-Channel Coding (JSCC), achieved through unsupervised end-to-end learning to facilitate image semantics transmission and reconstruction. This method effectively mitigates the inherent cliff effect in traditional communications and reduces bandwidth consumption. To bolster model adaptability across diverse Signal-to-Noise Ratio (SNR) environments, some models integrate an attention mechanism, incorporating SNR information into the model. Some researchers posit that different images may necessitate varying bandwidths to maintain consistent generation quality. Consequently, they introduce additional decision networks to gauge the requisite number of bits based on image semantics and channel conditions. However, JSCC-based methods of this nature encounter notable image quality degradation at low Compression Ratios (CRs). To tackle this challenge, some researchers bolster model performance at low CRs by integrating high-level semantics. Nevertheless, this approach mandates artificial labels, necessitating supplementary manual information for model training.

To improve image quality under low CRs and SNR conditions, this paper proposes a semantic communication model that can self-supervise to extract and utilize the high-level semantics of an image, and replacing the original end-to-end training process with two sub-processes, called semantic extraction and semantic decoding. In the semantic extraction sub-process, high-level semantics are incorporated into unsupervised training to enhance model robustness against noise. By refining the training approach of the self-supervised Momentum Contrast (MoCo) model and integrating high-level semantic information with low-level semantics through the Variational Autoencoder (VAE) model's decoder, the paper introduces the Momentum Contrast Semantic Encoder (MoCoSE) model tailored for semantic communication. For the semantic decoding sub-process, to enhance the quality of generated images, this paper establishes a mapping relationship from high-level semantic features to pixel information of the image using the classifier-free diffusion model. Simulation results demonstrate that under low CRs and SNR conditions, the model proposed in this paper outperforms four state-of-the-art models in semantic information transmission and restoration, generating images with more promising visual quality.

This work was supported by National Natural Science Foundation of China (General Program No. 62172342, Young Scientists Fund No. 62202392), the Natural Science Foundation of Hebei Province (No. F2022105027), and the Fundamental Research Funds for the Central Universities, P. R. China.