潜特征引导的条件扩散用于高保真生成图像语义通信

Zehao Chen*, Xinfeng Wei*, Haonan Tong*, Zhaohui Yang†, and Changchuan Yin*
* Beijing Laboratory of Advanced Information Network, Beijing University of Posts and Telecommunications, Beijing, China
† College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China
Emails: {chenzhzhz, xinfengwei, hntong, and ccyin}@bupt.edu.cn, yang_zhaohui@zju.edu.cn

摘要—语义通信被提出并期望提高第六代（6G）网络中大量数据传输的效率和有效性。然而，现有的基于深度学习的联合源信道编码（DeepJSCC）图像语义通信方案主要集中在优化像素级指标，并忽视了人类感知要求，从而导致感知质量下降。为了解决这一问题，我们提出了一种面向潜在表示的图像语义通信（LRISC）系统，该系统传输用于生成具有语义一致性的图像的潜在语义特征，从而确保接收端的感知质量。具体来说，我们首先通过基于神经网络（NN）的非线性变换将源图像映射到高维语义空间中的潜在特征。随后，使用自适应编码长度的联合源信道编码（JSCC）方案对这些特征进行编码以高效地传输无线信道。在接收端，开发了一个条件扩散模型，利用接收到的潜在特征作为条件指导来引导逆扩散过程，逐步重建高保真度图像同时保持语义一致性。此外，我们引入了一种信道信号噪声比（SNR）自适应机制，允许一个模型适用于各种信道状态。实验表明，所提出的方法在学习感知图像块相似性（LPIPS）和对抗信道噪声的鲁棒性方面显著优于现有方法，与 DeepJSCC 相比平均 LPIPS 减少了 43.3%，同时保证了语义一致性。

***Index Terms*—图像语义通信，联合源信道编码，条件扩散模型，语义一致性，自适应编码。**

I. 介绍

随着第六代（6G）网络的发展，对大规模连接和低延迟数据传输的需求迅速上升 [1]。传统的基于比特流的方法在处理大量数据和不断增加的设备连接数方面遇到困难 [2]–[4]。为了解决这一问题，以传达潜在意义而非原始数据为目标的语义通信作为一种有前景的解决方案出现，可以减少传输冗余并适应动态网络环境。生成模型结合联合源信道编码（JSCC），通过实现语义提取和编码过程提高了传输效率和鲁棒性。尽管现有的生成式语义通信能够显著提高通信效率，但现有技术仍缺乏确保生成图像语义通信中语义一致性和适应性的设计。

通过有效捕捉数据分布的特征，生成模型可以显著减轻深度联合源信道编码（DeepJSCC）[5] 中的语义失真，在生成性语义通信 [5]–[7] 领域展示了突破性的潜力。在 [7] 中的工作首次通过精炼图像的关键语义特征实现了语义压缩，并将这些语义特征作为解码过程中的引导信号使用。此外，为了增强重建的感知相似性，在 [8] 中的研究提取了多模态特征并将它们与信道状态信息结合作为条件来引导扩散模型逐步将初始随机噪声转化为最终重构数据。在 [9] 中的工作通过分解图像为空间范围和零空间分量，将生成性语义通信整合进预训练的扩散模型中，仅通过 DeepJSCC 传输空间范围特征并在接收端利用扩散模型重建零空间分量。

尽管 [5]–[9] 中的方法取得了改进，它们通常将预训练的扩散模型视为独立模块，这对于需要生成图像与原始源图像之间保持一致性的图像传输任务来说并不理想。此外，这些方法经常需要额外的数据传输来引导反向扩散过程，从而增加了系统开销。另外，现有方法通常执行固定的数据速率，无法自适应地确定不同类型数据源的最佳编码率。

为解决上述问题，我们提出了一种利用潜在特征引导的条件扩散模型实现高保真生成图像语义通信方案。本工作的关键贡献如下：(1) 我们提出了以潜在表示为导向的图像语义通信系统（LRISC），该系统集成了非线性变换编码与条件扩散模型。LRISC 能够通过基于潜在特征的条件扩散模型实现对潜在特征传输的自适应速率分配和高保真图像重建。(2) 在 LRISC 的发送端，我们采用变分熵建模来估计潜在特征分布，优化了速率分配以提高编码增益。在 LRISC 的接收端，我们利用潜在特征引导图像反扩散过程，确保原始图像与重构图像之间的语义一致性。(3) 我们引入了一个信道 SNR 自适应机制，使得一个模型可以在各种信道状态下工作。(4) 我们为编解码器开发了一种面向感知的训练策略，在损失函数中加入感知失真项以提高接收端的主观图像质量。仿真结果表明，所提出的方法在带宽受限的情况下优于现有方法，实现了更好的语义一致性和感知质量。

本文的其余部分组织如下。系统模型在第 II 节中介绍。模型架构和训练策略在第 III 节中详细说明。第 IV 节展示了模拟结果和分析，以证明所提系统的有效性。最后，第 V 节对论文进行了总结。

II. 系统模型

在 LRISC 系统中，我们首先通过非线性变换将源图像 $\boldsymbol{x}$ 映射到潜在特征 $\boldsymbol{z}$，然后进行 JSCC [10]。在接收端，引入了一个条件扩散模型作为解码器 [11]，该模型利用接收到的潜在特征 $\hat{\boldsymbol{z}}$ 作为条件信息来引导去噪反向扩散过程，并迭代地重构图像。

A. 非线性变换基于的编码器

如图 1所示，在发射端，采用非线性解析变换 g_e 从源图像 $\boldsymbol{x} \in \mathbb{R}^{n\times n}$ 中提取潜在空间中的语义特征，表示为 $\boldsymbol{z} = g_e(\boldsymbol{x};\boldsymbol{\theta}_g)$。这个潜在特征 $\boldsymbol{z}$ 被输入到先验编码器 h_e 和 JSCC 编码器 f_e 中。JSCC 编码器将特征 $\boldsymbol{z}$ 转换为适合信道传输的符号向量，从而得到传输向量

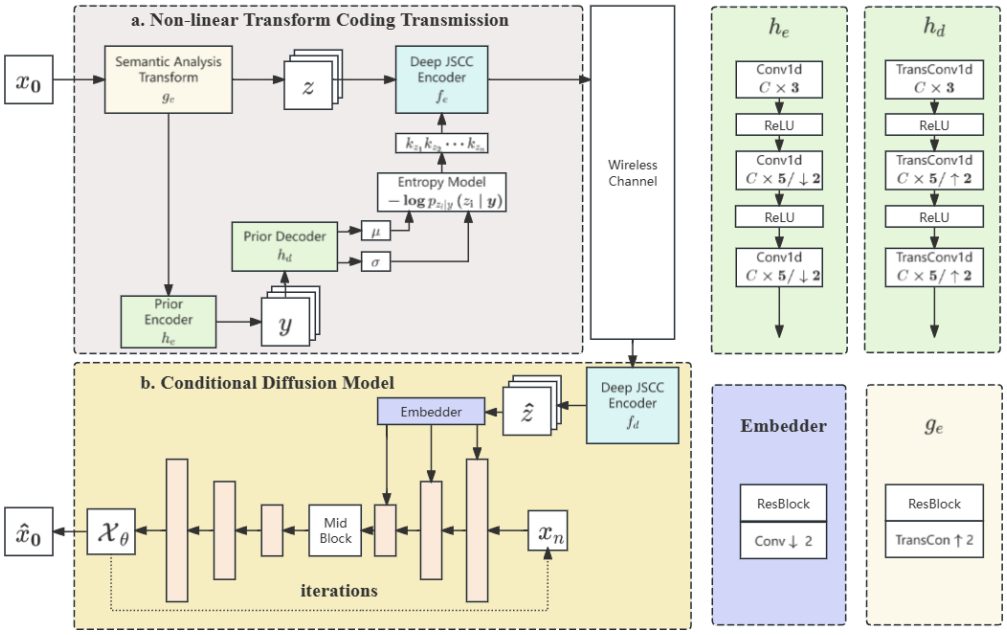

图 1. 网络的语义通信框架。

$\boldsymbol{s} = f_e(\boldsymbol{z};\boldsymbol{\theta}_f)$。给定信道传输函数 W，接收到的符号序列为 $\hat{\boldsymbol{s}} = W(\boldsymbol{s})$。具体来说，本文考虑加性白高斯噪声（AWGN）信道，因此接收端接收到的符号 $\hat{\boldsymbol{s}}$ 可表示为

$$\hat{\boldsymbol{s}} = W(\boldsymbol{s} \mid h) = h\boldsymbol{s} + \boldsymbol{n}, \tag{1}$$

其中，h 表示信道增益，$\boldsymbol{n} \sim \mathcal{N}(0, \sigma_n^2\, \boldsymbol{I})$ 是方差为 σ_n^2 的 AWGN，且 $\boldsymbol{I}$ 为单位矩阵。在接收端，接收到的符号 $\hat{\boldsymbol{s}}$ 被输入到 JSCC 解码器 f_d 中以重构潜在特征 $\hat{\boldsymbol{z}}$ 的估计值。重构的潜在特征 $\hat{\boldsymbol{z}} = f_d\left(\hat{\boldsymbol{s}};\boldsymbol{\phi}_f\right)$ 由 JSCC 解码器处理，其中 $\boldsymbol{\phi}_f$ 是 JSCC 解码器的可训练参数。然后利用条件扩散模型作为解码器 g_d 进行图像重构，得到 $\hat{\boldsymbol{x}}_0 = g_d\left(\hat{\boldsymbol{x}}_N, \hat{\boldsymbol{z}}; \boldsymbol{\phi}_g\right)$，其中 $\hat{\boldsymbol{x}}_N$ 是随机采样的高斯源。

一个先验编码器 h_e 捕获了 $\boldsymbol{z}$ 的全局依赖关系，并使用可学习的熵模型来表征速率分配的语义重要性，从而得到 $\boldsymbol{y} = h_e\left(\boldsymbol{z};\boldsymbol{\theta}_h\right)$，其中 $\boldsymbol{\theta}_h$ 表示熵编码的可训练参数。对于熵编码和解码，需要一个先验概率模型作为熵模型以确定每个符号的概率，确保更频繁出现的符号被赋予较短的码字。该熵模型如图 1 所示。潜特征 $\boldsymbol{z}$ 被建模为均值为 μ 和标准差为 σ 的多变量高斯变量向量。因此，$\boldsymbol{z}$ 的后验分布是

$$\begin{aligned} p_{\boldsymbol{z}|\boldsymbol{y}}(\boldsymbol{z} \mid \boldsymbol{y}) &= \prod_i \left(\mathcal{N}\left(z_i \mid \mu_i, \sigma_i\right) * \mathcal{U}\left(-\frac{1}{2}, \frac{1}{2}\right)\right)\left(z_i\right), \\ \text{with } (\boldsymbol{\mu}, \boldsymbol{\sigma}) &= h_d\left(\boldsymbol{z};\boldsymbol{\phi}_h\right), \end{aligned} \tag{2}$$

其中 $*$ 是卷积运算，而 $\mathcal{U}$ 是均匀分布。潜特征 $\boldsymbol{z}$ 包含多个嵌入向量 $\boldsymbol{z}_i$，其中 i 是第 i 维，每个的长度为 C。熵值 $\boldsymbol{z}_i$ 将被输入到 JSCC 编码器中，并将指导相应编码率的分配。当熵模型指示高熵时，编码器会为其传输分配更多资源。信道带宽成本 K_z 用于传输 $\boldsymbol{z}$ 可以由以下公式给出：

$$K_z = \sum_{i=1}^{k} \bar{k}_{\boldsymbol{z}_i} = \sum_{i=1}^{k} Q\left(k_{\boldsymbol{z}_i}\right) = \sum_{i=1}^{k} Q\left(-\beta \log p_{z_i|\boldsymbol{y}}\left(\boldsymbol{z}_i \mid \boldsymbol{y}\right)\right), \tag{3}$$

其中，β 是一个超参数，用于平衡从估计熵得出的信道带宽成本，$\bar{k}_{\boldsymbol{z}_i}$ 是 $\boldsymbol{z}_i$ 的带宽消耗，Q 表示一种 2^n 级标量量化，其量化值集设为 $\mathcal{V} = \{v_1, v_2, \ldots, v_{2^n}\}$。JSCC 将潜在特征 $\boldsymbol{z}$ 映射到信道传输符号 $\boldsymbol{s} \in \mathbb{C}^k$ 的过程由下式给出：

$$\boldsymbol{s} = f_e\left(\boldsymbol{z}, p_{\boldsymbol{z}_i|\boldsymbol{y}}\left(\boldsymbol{z}_i \mid \boldsymbol{y}\right); \boldsymbol{\theta}_{\boldsymbol{f}}\right). \tag{4}$$

为了满足实际通信系统的能量约束，信道传输符号在传输前必须满足平均功率约束，给定为 $\frac{1}{k}\sum_{i=1}^{k} |\boldsymbol{s}_i|^2 \leq P$ 其中 P 是最大传输功率。

B. 基于条件扩散模型的解码器

在接收端，我们采用条件扩散模型进行图像重建，潜在特征 $\hat{\boldsymbol{z}}$ 作为逆过程的条件。扩散过程（记为 q）逐步向图像 $\boldsymbol{x}_0$ 添加高斯噪声，形成一系列带噪数据 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$。逆扩散过程（记为 q_θ）学习逐步去噪。在步骤 n 中，两个马尔可夫过程 q 和 p_θ 可以分别描述为

$$q\left(\boldsymbol{x}_n \mid \boldsymbol{x}_{n-1}\right) = \mathcal{N}\left(\boldsymbol{x}_n | \sqrt{1-\beta_n}\boldsymbol{x}_{n-1}, \beta_n \boldsymbol{I}\right), \tag{5}$$

$$p_\theta\left(\boldsymbol{x}_{n-1} \mid \boldsymbol{x}_n, \hat{\boldsymbol{z}}\right) = \mathcal{N}\left(\boldsymbol{x}_{n-1} \mid M_\theta\left(\boldsymbol{x}_n, \hat{\boldsymbol{z}}, n\right), \beta_n \boldsymbol{I}\right). \tag{6}$$

其中变量 β_n 作为常数超参数。逆过程由神经网络（NN）$M_\theta\left(\boldsymbol{x}_n, \hat{\boldsymbol{z}}, n\right)$ 进行参数化。我们使用像素空间预测神经

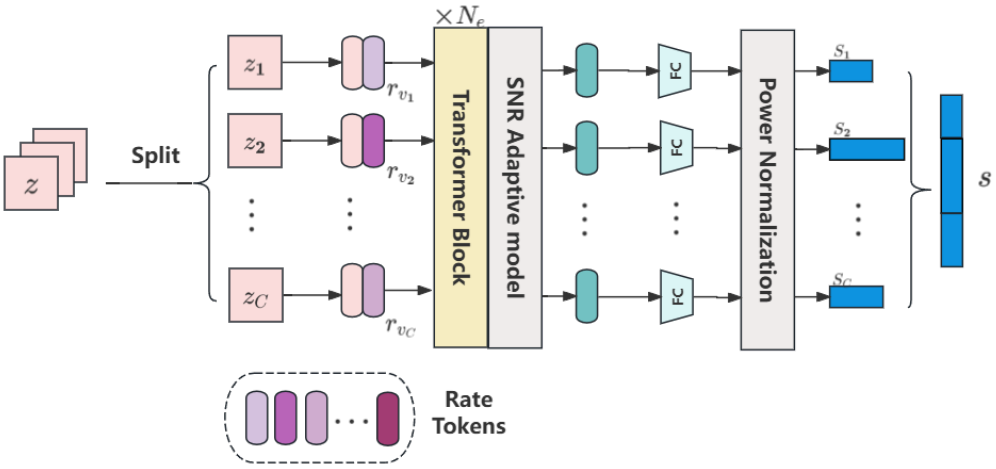

图 2. JSCC 编码器的架构。

网络 $\mathcal{X}_\theta$ 直接重构图像 x_0 而不是噪声 ϵ [12]。损失函数可以描述如下

$$\mathcal{L}(\theta, x_0) = \mathbb{E}_{\boldsymbol{x}_0,n,\epsilon} \frac{\alpha_n}{1-\alpha_n} \left\| \boldsymbol{x}_0 - \mathcal{X}_\theta \left(\boldsymbol{x}_n, \hat{\boldsymbol{z}}, \frac{n}{N_{\text{train}}} \right) \right\|^2, \tag{7}$$

其中 $n \sim \text{Unif}\{1, \ldots, N\}$，Unif 表示集合 $1, 2, \ldots, N$ 上的均匀分布，$\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 和 $\alpha_n = \prod_{i=1}^{n}(1-\beta_i)$。我们将模型条件化到伪连续变量 $\frac{n}{N_{\text{train}}}$ 上，这在选择解码的去噪步骤数量时提供了额外的灵活性。

模型训练完成后，可以通过遵循确定性生成的去噪扩散隐式模型（DDIM）[13] 生成图像，该模型是

$$\boldsymbol{x}_{n-1} = \sqrt{\alpha_{n-1}} \mathcal{X}_\theta \left(\boldsymbol{x}_n, \hat{\boldsymbol{z}}, \frac{n}{N} \right) + \sqrt{1-\alpha_{n-1}} \epsilon_\theta \left(\boldsymbol{x}_n, \hat{\boldsymbol{z}}, \frac{n}{N} \right). \tag{8}$$

III. 模型架构和训练策略

本节介绍了三个部分：LRISC 的神经网络架构、系统的损失函数以及多阶段训练算法。

A. 模型架构

1) **先前编解码器** h_e, h_d: h_e 和 h_d 的结构如图 1所示。它由全卷积层和 ReLU 激活函数组成。
2) ***JSCC* 编码器** $f_e f_d$: 我们采用了一对类似于变压器的 JSCC 编码器和解码器作为 f_e,f_d。编码器由 N_e 个变压器块和全连接（FC）层组成，以实现速率分配，如图 2所示。具体来说，$\boldsymbol{z}$ 首先被分离成嵌入序列 $z_1, z_2, \cdots, z_c$。由熵模型 $-\log P_{z_i|\boldsymbol{y}}(z_i \mid \boldsymbol{y})$ 引导，开发了一组可学习的速率标记嵌入，其维度与 z_i 相同 [10]，其中每一个对应于 $\mathcal{V} = \{v_1, v_2, \ldots, v_{2^n}\}$ 中的一个值，每个 z_i 与其对应的速率标记 r_{vi} 融合。全连接层的输出维度为 v_q 和 $q = 1, 2, \ldots, 2^n$，用于将嵌入映射到给定维度的 s_i。
3) **信噪比自适应模型**信噪比自适应模型被插入到 JSCC 编码器的最后一层和 JSCC 解码器的第一层，如图 3所示。它由 8 个全连接层与 7 个信噪比调制 (SM) 模块交替组成 [14]。SM 模块是一个三层的全连接网络，它将接收到的 z_i 的信噪比（记为 SNR_i）作为输入，并将其转换成一个 c 维的向量 sm_i。多个 SM 模块以粗到细的方式顺序级联，前面调制过的特征被送入后续的 SM 模块。
4) **条件扩散模型**去噪模块使用了 U-Net 结构 [15], [16]。每个 U-Net 单元由两个 ResNet 块、一个注意力块和一个卷积上采样及下采样块组成。条件扩散模型网络在下采样路径和上采样路径中均采用了六个 U-Net 单元。在下采样阶段，通道维度根据公式 $64 \times j$ 扩展，其中 j 表示层索引 $(j = 1, 2, ..., 6)$。上采样单元则按照相反顺序排列。对于嵌入条件 $\boldsymbol{z}$，我们采用 ResNet 块和转置卷积将 $\boldsymbol{z}$ 放大以匹配初始四个 U-Net 下采样单元的输入维度。

B. 损失函数

优化目标是在压缩性能和图像重建质量之间取得平衡，这可以使用失真加权损失函数表示如下：

$$\mathcal{L}_{RD} = \mathbb{E}_{\boldsymbol{x}_0} \left[-\lambda \log p_{\boldsymbol{z}|y}(\boldsymbol{z} \mid \boldsymbol{y}) - \lambda \log p_y(\boldsymbol{y}) + d(\boldsymbol{x}_0, \hat{\boldsymbol{x}_0}) \right], \tag{9}$$

其中拉格朗日乘子 λ 是一个权重因子，用于调整速率与失真的折衷关系，$d(\cdot, \cdot)$ 衡量原始图像和重建图像之间的失真程度，通过使用欧几里得距离（L2 范数）计算。为了符合人类感知质量，采用了一个额外的感知损失，表达如下：

$$\mathcal{L}_P = \mathbb{E}_{\boldsymbol{x}_0} \left[d_p \left(x_0, \hat{\boldsymbol{x}}_0 \right) \right], \tag{10}$$

其中 $d_p(\cdot, \cdot)$ 代表感知损失项。我们采用 LPIPS [17] 作为感知损失。为了提高训练的收敛性，我们在损失函数中添加了压缩失真项，率-失真-感知（RDP）损失函数定义为

$$\begin{aligned} \mathcal{L}_{\mathcal{RDP}} = \mathbb{E}_{\boldsymbol{x}_0} [& (1-\eta)[d(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0) + d(\boldsymbol{x}_0, \overline{\boldsymbol{x}}_0)] \\ & + \eta [d_p(\boldsymbol{x}_0, \hat{\boldsymbol{x}}_0) + d_p(\boldsymbol{x}_0, \overline{\boldsymbol{x}}_0)] \\ & + \lambda [-\log p_{\boldsymbol{z}|y}(\boldsymbol{z} \mid \boldsymbol{y}) - \log p_y(\boldsymbol{y})]], \end{aligned} \tag{11}$$

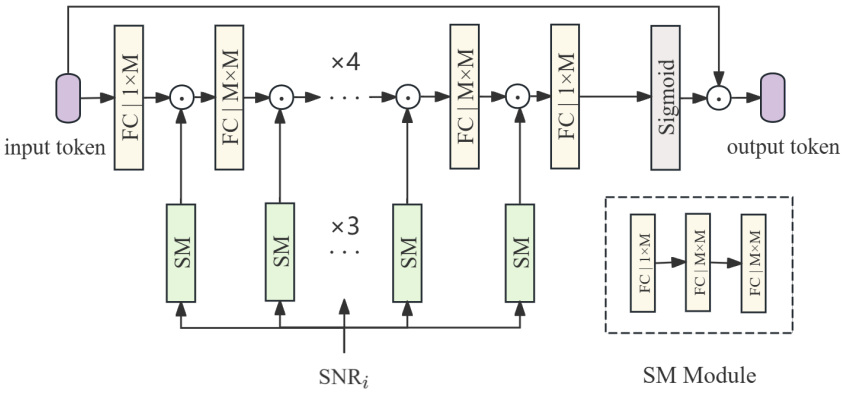

图 3. 信噪比自适应模型的架构。FC 是全连接网络，“⊙”是逐元素乘积。

Algorithm 1 提出的 LRISC 训练程序

Input: 训练数据集 X，拉格朗日乘子 λ 在速率项上，感知与失真权衡参数 η，从熵到信道带宽成本的缩放因子 β 和学习率 l_r。

阶段 1：训练 g_e, g_d, h_e, h_d
随机初始化所有参数并冻结参数 f_e 和 f_d
for training iteration 1 to N_{train} **do**
  采样 $\boldsymbol{x} \sim p_{\boldsymbol{X}}$,$n \sim \mathcal{U}(0, 1, 2, .., N_{\text{train}})$,$\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  $\overline{\boldsymbol{x}}_n = \sqrt{\alpha_n}\boldsymbol{x}_0 + \sqrt{1-\alpha_n}\epsilon)$
  $\boldsymbol{z} = g_e(\boldsymbol{x}_0) + \mathcal{U}(-0.5, 0.5)$
  $\overline{\boldsymbol{x}}_0 = \mathcal{X}_\theta(\overline{\boldsymbol{x}}_n, n/N_{\text{train}}, \boldsymbol{z})$
  $\mathcal{L}_{\mathcal{RDP}} = \mathbb{E}_{\boldsymbol{x}_0}[(1-\eta)d(\boldsymbol{x}_0, \overline{\boldsymbol{x}}_0)] + \eta[d_p(\boldsymbol{x}_0, \overline{\boldsymbol{x}}_0)]$
    $+\lambda[-\log p_{\boldsymbol{z}|y}(\boldsymbol{z} \mid \boldsymbol{y}) - \log p_y(\boldsymbol{y})]$
  更新参数 $(\boldsymbol{\theta}_g, \boldsymbol{\phi}_g, \boldsymbol{\theta}_h, \boldsymbol{\phi}_h)$
end for
阶段 2：训练 f_e, f_d
加载并冻结阶段 1 中训练的参数，并随机初始化 f_e 和 f_d
for each epoch **do**
  采样 $\boldsymbol{x} \sim p_{\boldsymbol{X}}$
  根据方程 (11) 计算损失函数
  更新参数：$(\boldsymbol{\theta}_f, \phi_f)$
end for
阶段 3：微调整个模型
加载前几个阶段训练的参数
for each epoch **do**
  重复步骤 14 到 16
  更新参数 $(\boldsymbol{\theta}_g, \boldsymbol{\theta}_h, \phi_g, \phi_h, \boldsymbol{\theta}_f, \phi_f)$
end for

Output: 参数 $(\boldsymbol{\theta}_g^*, \phi_g^*, \boldsymbol{\theta}_h^*, \phi_h^*, \boldsymbol{\theta}_f^*, \phi_f^*)$

其中 $\eta \in [0,1]$ 控制均方误差（MSE）**d**istortion 和 **p**erceptual loss 之间的权衡，而 λ 调整总传输 **r**ate 与重建质量之间的权衡。

C. 训练策略

为了确保训练的稳定性并提升整体性能，我们提出了一种多阶段训练策略。初始阶段，我们将每个模块单独训练 [18] 以降低复杂性并促进收敛 [19] 。当所有模块分别训练完成后，我们对整个模型进行微调。完整的多阶段训练过程如算法 1 所示。首先，训练 g_e,g_d,h_e 和 h_d。其次，使用信道数据训练 JSCC 编解码器 f_e 和 f_d。最后，以端到端优化的方式微调上述两个步骤中的所有模块。

IV. 仿真分析与讨论

A. 模拟设置

对于我们的实验，我们使用 COCO 数据集进行训练，并选择 50,000 张图像，这些图像是随机裁剪为 256×256 的尺寸，以及 Kodak 数据集用于测试。我们使用 PyTorch 框架实现所提出的 LRISC 模型。我们利用初始学习率为 1×10^{-4} 的 Adam 优化器，并将批量大小设置为 16。在第一阶段，我们随机初始化模型参数并训练 g_e、g_d、h_e 和 h_d，训练步骤设置为 10^6。在第二阶段，我们首先使用固定信道状态（SNR = 10dB）训练 JSCC 编解码器 f_e 和 f_d，除了 snr 自适应模块。然后，整个 JSCC 编解码器与 snr 自适应模块一起在可变信道状态下进行训练。最后，在微调过程中，所有模型参数都被解冻，并且该模型在完整的 COCO 训练数据集上再训练 20 个 epoch，学习率逐渐衰减。缩放因子 β 设置为 SNR = 10dB 时的 0.2，感知与失真权衡参数 η 设置为 0.5。训练过程在 NVIDIA RTX 4090 GPU 上进行以加速计算。

B. 性能比较

我们将所提出的方法与两种广泛认可的基线方法进行了比较：1）使用均方误差失真 [20] 优化的经典 DeepJSCC 模型，以及 2）传统的源通道分离方案。具体来说，源通道分离方案采用 BPG 进行源编码，并结合低密度奇偶校验 (LDPC) 代码进行信道编码，标记为“BPG + LDPC”。

对于性能评估，我们使用两个常用的指标：LPIPS 和峰值信噪比 (PSNR)。LPIPS 用于衡量感知相似度，因为它与人类视觉感知的相关性比传统的基于像素的指标如 PSNR 更高。而 PSNR 则提供了基于像素差异的重建质量的定量测量。

图 4比较了在不同信噪比水平下的性能。LRISC 在整个信噪比范围内始终达到最高的峰值信噪比。值得注意的是，在较低的信噪比值下，LRISC 和 DeepJSCC 展示出显著的鲁棒性，并且 LPIPS 相比于 DeepJSCC 减少了近 43.3%。这一改进突显了 LRISC 的优化联合源通

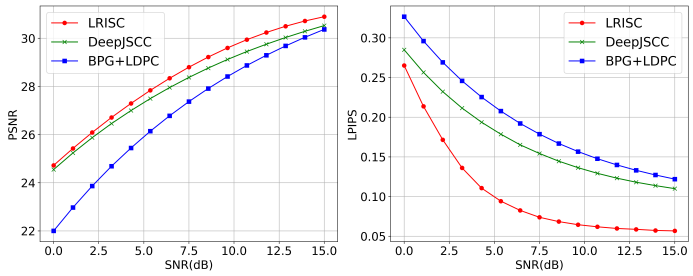

图 4. 峰值信噪比和感知指标性能对比信号噪声比。

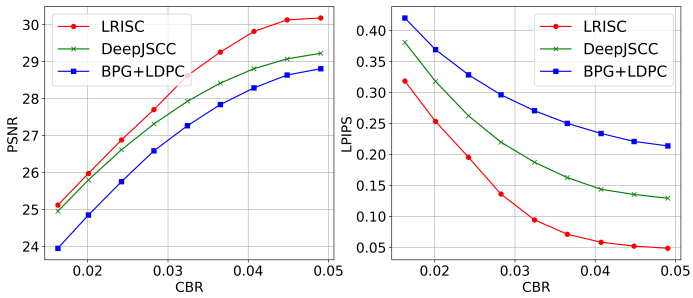

图 5. 峰值信噪比和感知指标性能对比恒定比特率。

道编码策略在噪声信道条件下保持图像质量的有效性。当考察 LPIPS 时，LRISC 和基线方法之间的性能差距随着信噪比的增加而扩大。这种改善是因为 LRISC 生成了更接近原始图像语义的感知信息。

图 5展示了不同方法在信道带宽比 (CBR) 变化时的性能对比。具体来说，LRISC 在 PSNR 方面始终优于 DeepJSCC 和 “BPG+LDPC” 基线。这一改进归因于我们 LRISC 中的增强压缩和自适应编码，这使得能够在更低的 CBR 下实现更高的重建质量。此外，LRISC 达到了最低的 LPIPS 值，表明其感知保真度优于其他方法。相比之下，BPG+LDPC 基线表现出最高的 LPIPS 值，表明尽管它在像素级质量方面保持可接受水平，但其感知质量仍然次优。

图 6展示了我们提出的框架中不同信噪比值下图像传输质量的视觉比较。从图 6可以看出，在最低信噪比水平（即，0 dB）时，重建的图像遭受严重的质量退化，表现出显著的视觉失真，但仍避免了传统方法中的 “悬崖效应” [21]。当信噪比增加到 5 dB 时，这些失真大大减轻。当信噪比达到 10 dB 或更高时，重建的图像几乎与原始源具有相同的保真度。这是因为传输的潜在特征有效地捕获了语义信息，从而在低信噪比条件下保持了强大的语义一致性，并显著提高了重建图像的感知质量。

V. 结论

在本文中，我们提出了一种由潜在特征驱动的生成式图像语义通信系统 LRISC，该系统专注于优化图像传输过程中的感知质量。具体来说，该系统通过使用熵模型实现潜变量的可变速率传输，并利用基于 Swin Transformer 的 JSCC 来减少通信资源消耗，同时接收方利用预训练扩散模型的强大生成能力，在有限带宽下实现高质量的图像重构。仿真结果表明，该系统不仅在语义图像传输中保持了鲁棒性，还提供了优越的感知质量。

参考文献

[1] P. Zhang, W. Xu, H. Gao, K. Niu, X. Xu, X. Qin, C. Yuan, Z. Qin, H. Zhao, J. Wei *et al.*, “Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks,” *Engineering*, vol. 8, pp. 60–73, 2022.

[2] S. Barbarossa, D. Comminiello, E. Grassucci, F. Pezone, S. Sardellitti, and P. Di Lorenzo, “Semantic communications based on adaptive generative models and information bottleneck,” *IEEE Communications Magazine*, vol. 61, no. 11, pp. 36–41, 2023.

[3] J. Dai, P. Zhang, K. Niu, S. Wang, Z. Si, and X. Qin, “Communication beyond transmitting bits semantics-guided source and channel coding,” *IEEE Wireless Communications*, vol. 30, no. 4, pp. 170–177, 2022.

[4] E. Grassucci, Y. Mitsufuji, P. Zhang, and D. Comminiello, “Enhancing semantic communication with deep generative models: An overview,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 13 021–13 025.

[5] E. Erdemir, T.-Y. Tung, P. L. Dragotti, and D. Gündüz, “Generative joint source-channel coding for semantic image transmission,” *IEEE Journal on Selected Areas in Commun.*, vol. 41, no. 8, pp. 2645–2657, 2023.

[6] T. Wu, Z. Chen, D. He, L. Qian, Y. Xu, M. Tao, and W. Zhang, “Cddm: Channel denoising diffusion models for wireless semantic communications,” *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 11 168–11 183, 2024.

[7] L. Qiao, M. B. Mashhadi, Z. Gao, C. H. Foh, P. Xiao, and M. Bennis, “Latency-aware generative semantic communications with pre-trained diffusion models,” *IEEE Wireless Communications Letters*, vol. 13, no. 10, pp. 2652–2656, 2024.

[8] E. Lei, Y. B. Uslu, H. Hassani, and S. S. Bidokhti, “Text+ sketch: Image compression at ultra low rates,” *arXiv preprint arXiv:2307.01944*, 2023.

[9] M. Yang, B. Liu, B. Wang, and H.-S. Kim, “Diffusion-aided joint source channel coding for high realism wireless image transmission,” *arXiv preprint arXiv:2404.17736*, 2024.

[10] J. Dai, S. Wang, K. Tan, Z. Si, X. Qin, K. Niu, and P. Zhang, “Nonlinear transform source-channel coding for semantic communications,” *IEEE Journ. on Select. Areas in Commun.*, vol. 40, no. 8, pp. 2300–2316, 2022.

[11] R. Yang and S. Mandt, “Lossy image compression with conditional diffusion models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 971–64 995, 2023.

[12] T. Salimans and J. Ho, “Progressive distillation for fast sampling of diffusion models,” *arXiv preprint arXiv:2202.00512*, 2022.

[13] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *arXiv preprint arXiv:2010.02502*, 2020.

[14] K. Yang, S. Wang, J. Dai, X. Qin, K. Niu, and P. Zhang, “Swinjscc: Taming swin transformer for deep joint source-channel coding,” *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[15] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer International Publishing, 2015, pp. 234–241.

[16] Y. Zeng, X. He, X. Chen, H. Tong, Z. Yang, Y. Guo, and J. Hao, “DMCE: Diffusion model channel enhancer for multi-user semantic communication systems,” in *IEEE International Conf. on Commun.*, 2024, pp. 855–860.

[17] R. Zhang, P. Isola, and A. Efros, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[18] X. Wei, H. Tong, N. Yang, and C. Yin, “Language-oriented semantic communication for image transmission with fine-tuned diffusion model,” in *Wireless Commun. and Signal Processing (WCSP)*, Oct. 2024.

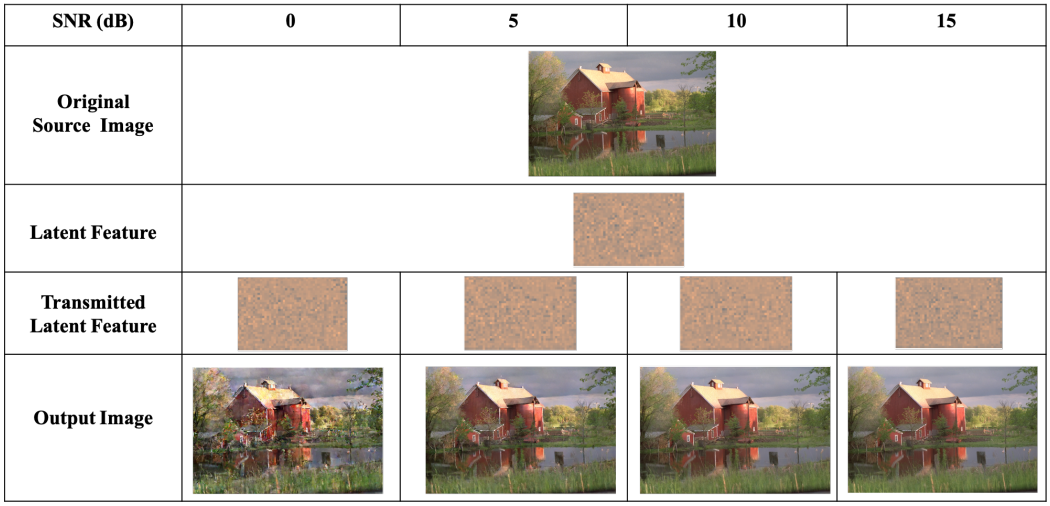

图 6. 所提框架的可视化示例。

[19] H. Tong, H. Li, H. Du, Z. Yang, C. Yin, and D. Niyato, "Multimodal semantic communication for generative audio-driven video conferencing," *IEEE Wireless Communications Letters*, vol. 14, no. 1, pp. 93–97, 2025.

[20] E. Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 4774–4778.

[21] Q. Pan, H. Tong, J. Lv, T. Luo, Z. Zhang, C. Yin, and J. Li, "Image segmentation semantic communication over internet of vehicles," in *2023 IEEE Wireless Commun. and Net. Conf. (WCNC)*, 2023.