http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2024.0542

基于 Swin Transformer 的模分语义图像传输

彭艺[1,2]，孙畅[1,2]，杨青青[2,*]，李辉[2]，王健明[2]

(1. 昆明理工大学 信息工程与自动化学院，昆明 650031；2. 昆明理工大学 云南省计算机技术应用重点实验室，昆明 650500)

摘 要：语义通信作为下一代通信系统的新范式，极大提高了通信效率。然而，此前的语义图像传输工作大多是建立在卷积神经网络 (CNN) 上进行的，传统的 CNN 模型通常难以全面挖掘图像中的语义信息。为解决该问题，采用 Swin Transformer 作为骨干网络来提取语义信息。在此基础上，提出基于模分的语义图像传输新方法，通过特征分离实现公共特征和个性特征的独立传输。在输入的图像经过联合信源信道编码 (JSCC) 后，对所编码后的语义特征进行公共特征与个性特征的区分并分别进行传输，通过对公共特征的资源共享，有效降低了带宽需求，从而提升了传输效率。为评估模型性能，在公开的 CIFAR10 数据集上，将所提方法与传统方法进行对比，实验结果表明：所提方法有效提高了图像传输质量，并在传输效率和带宽节省方面取得了良好的性能。

关 键 词：语义通信；图像传输；模分多址；图像重建；Swin Transformer

中图分类号：TN929.5; TP391.41; TP18

文献标志码：A **文章编号**：1001-5965(2026)11-3842-10

随着通信技术的进步，传统通信系统在信息压缩、传输和处理上遇到瓶颈，迫切需要创新[1]。新兴应用(如自动驾驶和物联网)对通信提出更高要求，不仅需准确传输比特，还要支持智能决策和人机交互[2]，促使从业者深入探讨全新设计思路。语义通信因其与人工智能的紧密结合[3]及在提高通信效率方面的巨大潜力，成为研究热点[4]。

与经典香农信息论不同，语义通信关注信息的语义特征，并通过相应机制进行传输，能够在带宽受限的情况下超越香农容量极限，从而突破香农极限及冯诺依曼架构的限制[5]。香农在 1948 年提出的经典信息论关注的是符号的可靠传输，而未考虑更高层次的语义层面设计。香农的忽视源自当时的工程需求，但并不意味着应忽略信息的意义[6]。

得益于人工智能在算法、算力、数据方面的巨大进步，最近涌现了大量端到端语义通信系统模型，包括面向文本的语义通信[7]、面向图像的语义通信[8]及面向语音的语义通信[9]。这些语义通信系统类似于自动编码器，通过压缩和重构信号来优化传输并保留其语义信息。但其区别于去噪自编码器，在去噪自编码器中原始图像已经包含伪影，而在语义通信系统中则是将扰动添加到已经压缩的潜在特征上。这些扰动可以理解为在信号传输过程中由于无线信道所造成的干扰，如瑞利衰落或加性高斯白噪声(additive white Gaussian noise, AWGN)[10]。

近年来，深度学习模型在分析文本和图像等高维数据时展现了出色的能力。基于深度学习的联合信源信道编码(deep joint source-channel coding, Deep JSCC)[11]及其衍生版本[12-13]主要采用卷积神经网络(convolutional neural networks, CNN)作为骨干网络，直接将图像像素映射到信道输入符号，从而提升无线图像传输的效率和性能，超越了传统分离式方法(如 JPEG/BPG)。文献 [14] 提出一种基于注意

收稿日期：2024-07-18；**录用日期**：2024-12-20；**网络出版时间**：2025-01-16 13:40
网络出版地址：link.cnki.net/urlid/11.2625.V.20250115.1641.001
基金项目：国家自然科学基金 (62461030)；云南省基础研究计划重点项目 (202401AS070105)
* **通信作者**. E-mail：12309214@kust.edu.cn

引用格式：彭艺，孙畅，杨青青，等. 基于 Swin Transformer 的模分语义图像传输 [J]. 北京航空航天大学学报，2026，52（11）：3842-3851.
Peng Y，Sun C，Yang Q Q，et al. Swin Transformer-based semantic image transmission with model division multiplexing[J]. Journal of Beijing University of Aeronautics and Astronautics，2026，52（11）：3842-3851 （in Chinese）.

力机制的深度学习联合信源信道编码(attention-based deep joint source-channel coding, ADJSCC)方法, 通过信道级软注意力机制动态调整特征缩放, 解决了不同信噪比(signal-to-noise ratio, SNR)条件下的适应性和鲁棒性问题。

然而, 传统 CNN 模型通常受到局部感受野和层级限制的影响, 使其难以全面挖掘图像中的语义信息。国内外研究者受到 Vision Transformer (ViT)在计算机视觉领域中最新进展的启发, 开始将其作为骨干网络应用于语义通信领域[15-17]。相比于传统 CNN 模型, Swin Transformer 的窗口注意力机制通过减少全局计算的开销提升了计算效率。另外, 与 ViT 相比, Swin Transformer 采用局部窗口注意力机制, 使其计算复杂度与图像大小呈线性关系, 而传统 ViT 模型则表现出二次复杂度[15]。文献 [16] 以 Swin Transformer 作为骨干网络进行信息提取, 并在此基础上提出了空间调制模块提升信道适应性, 与 ADJSCC 方法相比, 高效的窗口注意力机制将浮点运算次数成本降低了 2.5 倍, 尽管模型较大, 但提供了更优的性能和更快的速度。文献 [17] 提出一种基于 Swin Transformer 的多用户语义通信系统, 通过自适应调整传输信号长度提升图像重建质量并避免网络拥堵。文献 [18] 结合传统 CNN 和现代视觉模型的优势, 引入一种基于 ConvNext 的新型语义通信结构, 降低了计算开销, 提升了整体的传输效率。但现有基于 Swin Transformer 的语义图像传输系统研究主要集中在语义信息提取、信道适应性提升及数据压缩等方面。尽管这些方法有效提升了图像传输的性能, 但大多侧重于提升传输的鲁棒性或压缩比, 未能充分考虑如何优化带宽需求。

为应对高维信息空间所带来的挑战, 引入了简化信息传输的方法, 旨在通过优化信息原语层面减少冗余, 提高无线网络的处理效率。根据文献 [19-20], 从相同语义通信模型中提取的语义特征包含一些共性的语义信息, 这些信息被称为公共特征, 其余部分则被称为个性特征。受此启发, 本文在图像传输系统中加入公共特征提取器, 来对提取出来的语义特征进行区分, 从而达到减少传输过程中的冗余信息、提高传输效率的目的。

基于此, 本文提出一种面向图像传输的语义通信系统框架。通过引入 Swin Transformer 网络架构及公共特征提取器, 克服了传统通信系统的技术瓶颈, 减少了传输过程中的冗余信息, 并有效节约了带宽资源, 从而提升了端到端传输性能。

1　语义通信系统

近年来, 随着深度学习的快速发展和广泛应用, 基于深度学习的语义编解码技术成为一个非常有潜力的重要发展方向。在语义通信系统中运用深度学习算法, 可以打破语义提取缺乏显式数学表达的障碍, 使系统更加灵活。

图 1 展示了一个传统的数字通信系统。在该系统中, 信源通过压缩和编码模块生成比特流, 并通过无线信道进行传输。接收端根据解码过程恢复数据, 并重构原始信息。

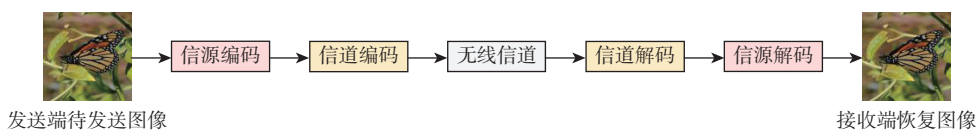


图 1　传统数字通信系统

Fig. 1　Traditional digital communication system

图 2 和图 3 展示了目前最常见的 2 种基于神经网络的语义通信系统。图 2 是一种较典型的神经网络语义通信系统, 对于现有通信系统改变是最小的, 其首先利用语义提取神经网络对原始数据进行预处理, 随后将提取出来的语义信息分别通过信源编码和信道编码。而图 3 是另一种基于神经网络的语义通信系统, 其利用神经网络对语义提取、信源编码及信道编码进行联合设计, 实现端到端的系统框架。其优点在于, 利用了在码长有限长的情况下联合信源信道编码(joint source-channel coding, JSCC)优于分离编码这一特点。

香农分离定理表明, 在传输无限长比特时, 独立的信源和信道编码可以提供最佳性能, 但这一假设在实际中难以实现。因为信源编码和信道编码会相互影响接收到的信息, 导致这种理想化模型不完全适用。因此, JSCC 在语义通信中逐渐受到重视。随着深度学习技术的发展, Deep JSCC 方法的引入使联合编码和解码变得更加高效和可行。

端到端 JSCC 相比传统分离式编码的最大突破在于通过深度学习将编码器、解码器和无线信道一起端到端训练。以往模型需在特定 SNR 下训练, 并在预测阶段保持相似 SNR, 否则性能下降。近年来, 引入注意力机制并根据 SNR 调整信源语义特征权重的研究, 使单一模型能适应不同信道条件,

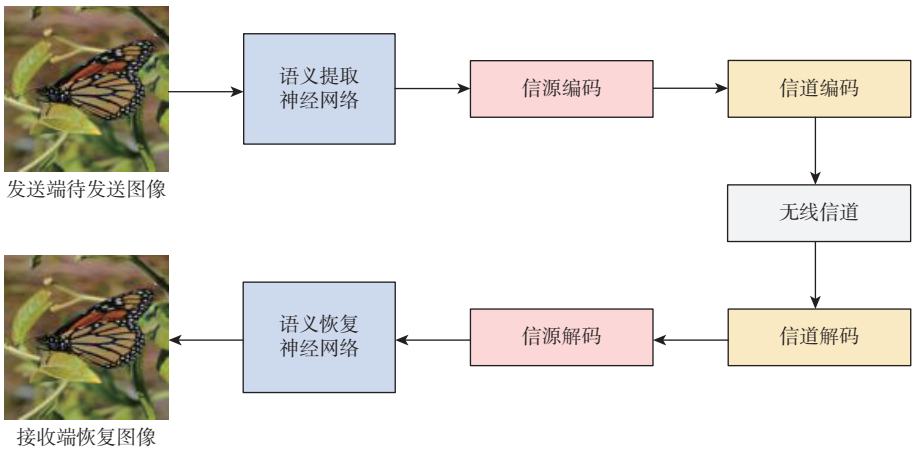


图 2 基于神经网络的语义通信系统 1

Fig. 2 Neural network-based semantic communication system 1



图 3 基于神经网络的语义通信系统 2

Fig. 3 Neural network-based semantic communication system 2

避免了多次训练。

2 模分图像语义传输方法

2.1 系统架构

图 4 展示了基于 Swin Transformer 的模分语义图像传输系统模型，其主要由发送端、通信信道和接收端 3 个部分组成。在发送端，将信源样本表示为 $\boldsymbol{I}\in\mathbf{R}^N$，其中，$N$为向量的长度，可以通过样本高度×样本宽度×样本通道数来进行计算。从$\boldsymbol{I}$中提取的语义特征向量表示为

$$\boldsymbol{S}_I = E_\alpha(\boldsymbol{I}) \tag{1}$$

式中：$E_\alpha(\cdot)$为语义编码器，用于提取信源样本的深层语义特征；α为语义特征提取函数中的可训练参数。

提取出来的语义特征向量输入信道自适应缩放模块中，根据通道状态将其进行重新缩放，从而增强在无线环境中传输符号的抗噪声和抗干扰能力。信道自适应缩放模块由多个全连接层和 SNR 调制（SNR modulation, SM）模块交替组成[16]，其可以根据信道状态对语义特征进行调制，以适应不同的信道状态。编码过程可以被表示为

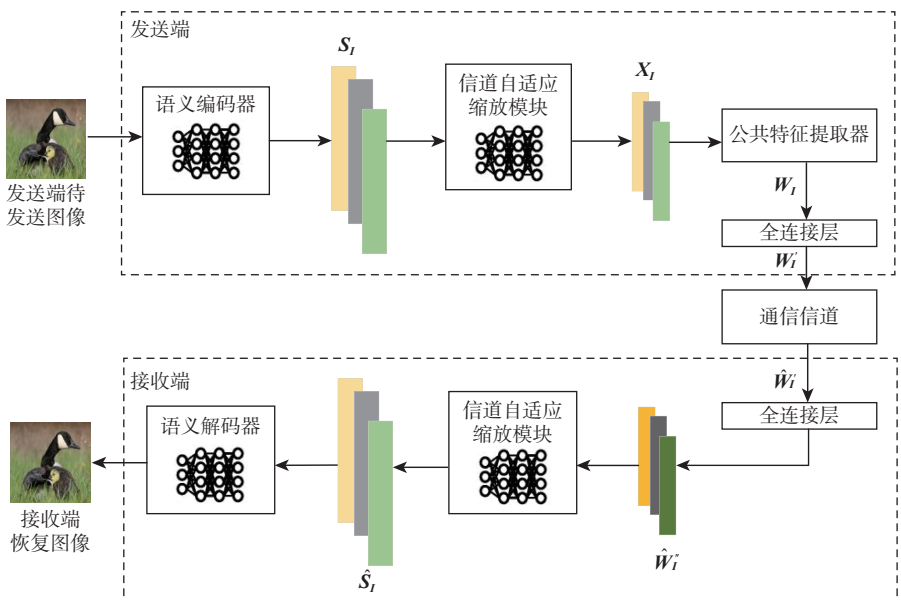


图 4 模分语义图像传输系统模型

Fig. 4 Semantic image transmission system model with model division multiplexing

$$\boldsymbol{X}_{\boldsymbol{I}} = C_{\beta}(\boldsymbol{S}_{\boldsymbol{I}}) \in \mathbf{R}^{k} \tag{2}$$

式中: $C_{\beta}(\cdot)$为具有参数β的信源信道编码函数; k为信道编码后的语义特征$\boldsymbol{X}_{\boldsymbol{I}}$的长度。

信源信息经过语义编码、信道编码后, 送入公共特征提取器中, 公共特征提取器可以对其进行区分, 具体来说, 公共特征提取器将这些语义特征区分为公共特征$\boldsymbol{W}_{\boldsymbol{I}\mathrm{c}}$与个性特征$\boldsymbol{W}_{\boldsymbol{I}\mathrm{i}}$, 该过程可以表述为

$$\boldsymbol{W}_{\boldsymbol{I}} = f_{\mathrm{c}}(\boldsymbol{X}_{\boldsymbol{I}}) = \{\boldsymbol{W}_{\boldsymbol{I}\mathrm{i}}, \boldsymbol{W}_{\boldsymbol{I}\mathrm{c}}\} \tag{3}$$

式中: $f_{\mathrm{c}}(\cdot)$为提取公共特征的函数; $\boldsymbol{W}_{\boldsymbol{I}}$为经过公共特征提取器后的总语义特征。

而后经历一个全连接层$M_{\mathrm{f}}(\cdot)$将语义特征压缩变成C维语义特征$\boldsymbol{W}_{\boldsymbol{I}}'$。信道带宽比(channel bandwidth ratio, CBR)定义为信道输入维度与信源维度的比值, 其直接影响图像信息的保留程度, 从而决定重建图像的质量。峰值信噪比(peak signal-to-noise ratio, PSNR)用于评估图像重建质量, 由于其值与重建误差成反比, 因此, CBR 的减小则导致信息丢失, 重建误差增加, 会显著降低 PSNR 性能。

经过公共特征提取器区分与压缩后的语义特征$\boldsymbol{W}_{\boldsymbol{I}}'$通过无线信道进行传输, 此处考虑 AWGN 信道, 并将信道建模为一个不可训练的神经网络层[11]。在传输信号$\boldsymbol{W}_{\boldsymbol{I}}'$过程中, AWGN 对信号产生干扰, 接收到的信号$\hat{\boldsymbol{W}}_{\boldsymbol{I}}'$由接收器获取, 可表示为

$$\hat{\boldsymbol{W}}_{\boldsymbol{I}}' = G\left(\boldsymbol{W}_{\boldsymbol{I}}'; \sigma^{2}\right) = \boldsymbol{W}_{\boldsymbol{I}}' + \boldsymbol{n} \tag{4}$$

式中: $G(\cdot)$为信道传输函数; 噪声分量$\boldsymbol{n}$服从高斯分布, $\boldsymbol{n} \sim N\left(0, \sigma^{2}\right)$; σ^{2}为噪声功率。

接收端主要包含信道自适应缩放模块及语义解码器, 两者分别对应发送端的自适应缩放过程和语义编码过程的逆过程。当接收端接收到信号$\hat{\boldsymbol{W}}_{\boldsymbol{I}}'$, 先经过一个全连接层$M_{\mathrm{f}}^{-1}(\cdot)$恢复为原维度语义特征$\hat{\boldsymbol{W}}_{\boldsymbol{I}}''$, 随后采用信道解码函数$C_{\phi}^{-1}(\cdot)$解码出语义特征$\hat{\boldsymbol{S}}_{\boldsymbol{I}}$, 再利用语义解码器$\boldsymbol{E}_{\eta}^{-1}(\cdot)$恢复出信源信息$\hat{\boldsymbol{I}}$。整个传输过程可表示为$\boldsymbol{I} \xrightarrow{E_{\alpha}(\cdot)} \boldsymbol{S}_{\boldsymbol{I}} \xrightarrow{C_{\beta}(\cdot)} \boldsymbol{X}_{\boldsymbol{I}} \xrightarrow{f_{\mathrm{c}}(\cdot)} \boldsymbol{W}_{\boldsymbol{I}} \xrightarrow{M_{\mathrm{f}}(\cdot)} \boldsymbol{W}_{\boldsymbol{I}}'$, $\boldsymbol{W}_{\boldsymbol{I}}' \xrightarrow{G(\cdot)} \hat{\boldsymbol{W}}_{\boldsymbol{I}}' \xrightarrow{M_{\mathrm{f}}^{-1}(\cdot)} \hat{\boldsymbol{W}}_{\boldsymbol{I}}'' \xrightarrow{C_{\phi}^{-1}(\cdot)} \hat{\boldsymbol{S}}_{\boldsymbol{I}} \xrightarrow{\boldsymbol{E}_{\eta}^{-1}(\cdot)} \hat{\boldsymbol{I}}$。

在此, 语义特征编、解码函数使用 Swin Transformer 构建, 整个系统能更快速地从带噪声信号中重建原始图像。该系统的训练损失函数为

$$\min_{\boldsymbol{\psi},\boldsymbol{\gamma}} \boldsymbol{E}_{\boldsymbol{I}\sim\boldsymbol{P}_{\boldsymbol{I}}} \boldsymbol{E}_{\hat{\boldsymbol{W}}_{\boldsymbol{I}}'\sim\boldsymbol{P}_{\hat{\boldsymbol{W}}_{\boldsymbol{I}}'|\boldsymbol{I}}} [d(\boldsymbol{I}, \hat{\boldsymbol{I}})] \tag{5}$$

式中: $\boldsymbol{\psi}$、$\boldsymbol{\gamma}$分别为编码器、解码器中所有的网络参数; $\boldsymbol{P}_{\boldsymbol{I}}$为与信源样本$\boldsymbol{I}$相关的概率分布; $\boldsymbol{P}_{\hat{\boldsymbol{W}}_{\boldsymbol{I}}'|\boldsymbol{I}}$为给定信源样本$\boldsymbol{I}$时, 关于接收符号序列$\hat{\boldsymbol{W}}_{\boldsymbol{I}}'$的条件概率分布; d为原始图像与重建图像之间的差异。

2.2　语义编、解码部分

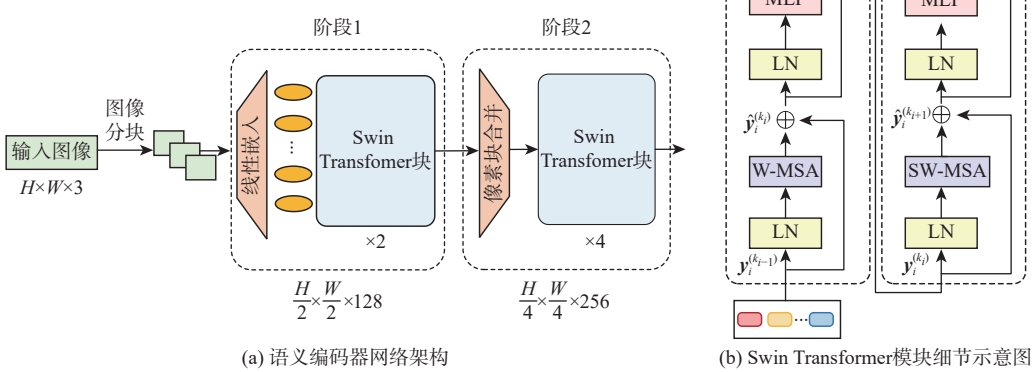
语义编码器的网络架构如图 5(a)所示, 采用 Swin Transformer 提取图像语义特征, 构成语义编码器的核心单元, 实现信源语义信息的处理。先将图像分割为若干不重叠的图像块, 经过线性嵌入后, 通过 2 个 Swin Transformer 块对这些图像块进行处理(阶段 1)。再将这些图像块送到阶段 2, 该阶段由像素块合并层及 4 个 Swin Transformer 块构成。语义解码器采用与语义编码器对称的结构。

如图 5(b)所示, Swin Transformer 块通过将多头自注意力(multi-head self-attention, MSA)模块替换为移位窗口模块, 实现了窗口分区的移动, 增强了全局信息提取能力, 并提高了特征提取效率。使用移动窗口分组方法后, 在第i个阶段中连续的 Swin Transformer 块k_i和k_{i+1}可以进行如下计算:

$$\hat{\boldsymbol{y}}_{i}^{(k_i)} = \text{W-MSA}\left(\text{LN}\left(\boldsymbol{y}_{i}^{(k_{i-1})}\right)\right) + \boldsymbol{y}_{i}^{(k_{i-1})} \tag{6}$$

$$\boldsymbol{y}_{i}^{(k_i)} = \text{MLP}\left(\text{LN}\left(\hat{\boldsymbol{y}}_{i}^{(k_i)}\right)\right) + \hat{\boldsymbol{y}}_{i}^{(k_i)} \tag{7}$$

$$\hat{\boldsymbol{y}}_{i}^{(k_{i+1})} = \text{SW-MSA}\left(\text{LN}\left(\boldsymbol{y}_{i}^{(k_{i+1})}\right)\right) + \boldsymbol{y}_{i}^{(k_{i+1})} \tag{8}$$



(a) 语义编码器网络架构　　(b) Swin Transformer模块细节示意图

图 5　语义编码器网络架构及 Swin Transformer 模块细节示意图[15]

Fig. 5　Semantic encoder architecture and Swin Transformer module details diagram[15]

$$\boldsymbol{y}_i^{(k_{i+1})}=\mathrm{MLP}\left(\mathrm{LN}\left(\hat{\boldsymbol{y}}_i^{(k_{i+1})}\right)\right)+\hat{\boldsymbol{y}}_i^{(k_i)} \tag{9}$$

式中：$\hat{\boldsymbol{y}}_i^{(k_i)}$和$\boldsymbol{y}_i^{(k_i)}$分别为第$i$个阶段的块$k_i$中(S)W-MSA模块和MLP模块的输出特征；W-MSA和SW-MSA为Swin Transformer中的MSA模块，具有常规和移动窗口配置；MLP为多层感知机，用于对特征作进一步映射和信息融合；LN表示层归一化操作。

随后，每个从语义编码器中输出的语义特征将根据信道状态通过信道编码器重新进行缩放。

2.3 信道自适应缩放模块

信道自适应缩放模块分别由具有不同全连接层的多层感知器构成，其可以根据不同的信道状态生成特定的调制策略，以适应信道的变化。图6展示了信道编码和解码网络架构，其由8个全连接层和7个SM模块交替组成。每个SM模块包含3层全连接网络，根据输入的信号S_m，将其映射为m维的向量$\boldsymbol{s}_m$，以调整隐性特征各通道之间的比例信息。SM模块通过动态调整隐性特征的比例，提升网络对不同SNR条件下的适应性和鲁棒性，从而优化编、解码性能。具体映射过程为

$$\boldsymbol{s}_m^{(1)}=\mathrm{ReLU}(\boldsymbol{A}^{(1)}S_m+\boldsymbol{b}^{(1)}) \tag{10}$$

$$\boldsymbol{s}_m^{(2)}=\mathrm{ReLU}(\boldsymbol{A}^{(2)}\boldsymbol{s}_m^{(1)}+\boldsymbol{b}^{(2)}) \tag{11}$$

$$\boldsymbol{s}_m=\mathrm{ReLU}(\boldsymbol{A}^{(3)}\boldsymbol{s}_m^{(2)}+\boldsymbol{b}^{(3)}) \tag{12}$$

式中：ReLU为激活函数；$\boldsymbol{A}^{(\cdot)}$和$\boldsymbol{b}^{(\cdot)}$分别为仿射函数的参数及其对应的偏移量。

此后，级联多个相同的SM模块，再经过一个Sigmod激活函数输出[0, 1]之间的权重值，以缩放Swin Transformer模块的输出特征向量。

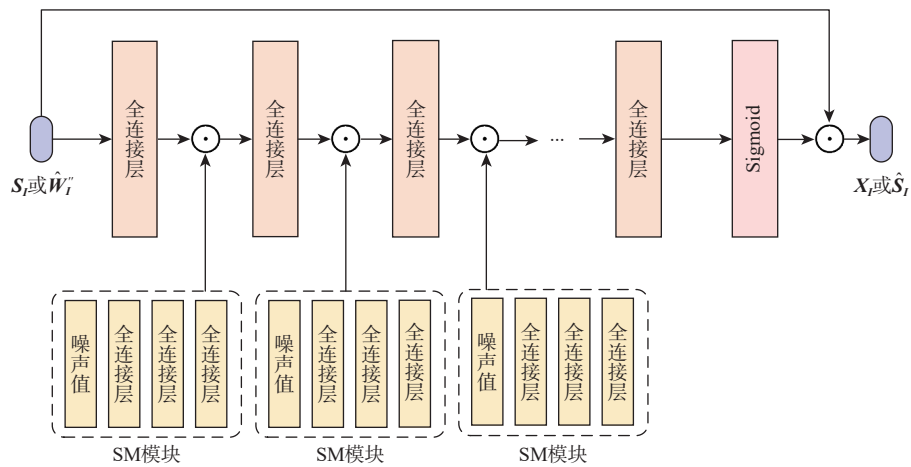


图6 信道自适应缩放模块网络架构[16]

Fig. 6 Channel adaptive scaling module network architecture[16]

2.4 公共特征提取方法

受到模分多址技术的启发，本文工作的目标是在2.1～2.3节所述通信系统模型的基础上，识别、分析、提取语义信息中所代表的共性和个性化信息，对携带共性信息的信号进行复用，以进一步降低传输带宽。根据文献[19]，从相同数据集的不同图像提取出来的语义信号非常相似，可以从高维语义空间中挖掘不同图像所传输语义信息的共性和个性信息，从而充分利用从高维语义空间中挖掘出来的共性信息，降低由冗余传输而造成的带宽消耗。

利用以上方法对CIFAR10数据集中同类别的任意2张图像提取语义信息。如图7所示，其表示从同类别图像中提取出的相关语义信息集合。可以看出，虽然是从不同图像中提取出的信号图像，但其语义信息集非常相似。

在公共特征提取器中，对提取出来的不同语义向量每个维度的方差进行计算。通过分析每个维度的方差，可以评估特征在不同维度上的离散程度。方差小的特征维度表示该维度的信息在多个样本中变化较小；相反，方差大的特征维度表示该维度的信息在样本之间变化较大。每个特征维度的方差可表示为

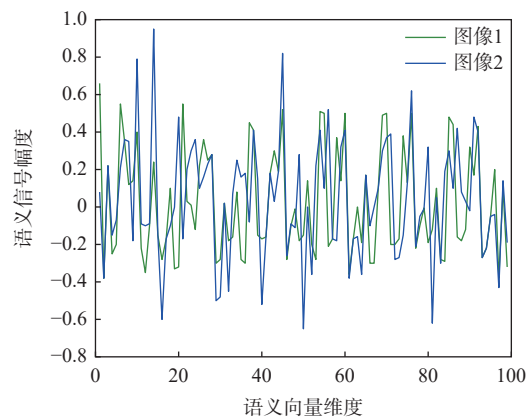


图7 语义特征变化趋势

Fig. 7 Trend of semantic feature variation

$$V_{\mathrm{ar}}(x_{i'}) = \frac{1}{N'}\sum_{j=1}^{N'}\left(x_{i'j}-\mu_{i'}\right)^2 \tag{13}$$

式中: $x_{i'j}$为第i'个特征维度在第j个样本中的值; $\mu_{i'}$为第i'个维度的均值; N'为样本数量。

为将提取出的语义特征区分为公共特征和个性特征, 设定阈值θ, 当某一特征维度的方差小于或等于θ时, 则被认为是公共特征; 当方差大于θ时, 则被认为是个性特征[19]。因此, θ越高, 更多特征会被判定为公共特征, 从而提高带宽效率, 这是因为较高的θ将允许变化较小的特征被视为图像间的共性, 但这可能导致忽略部分个性特征, 进而影响图像的质量。θ的选取可以通过实验调优, 以确保特征的有效分类。

从公共特征提取器中区分出的特征都会经过一个全连接层进行压缩, 将语义特征进行压缩, 进入通信信道中。全连接层的主要作用是对语义特征进行降维, 从而减少传输过程中所需的带宽, 同时, 保留关键的语义信息, 确保高质量的图像传输。具体来说, 假设输入的语义特征维度为M, 全连接层的输出为C维, 该压缩过程可表示为

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b} \tag{14}$$

式中: $\boldsymbol{W}\in\mathbf{R}^{C\times M}$为全连接层的权重矩阵; $\boldsymbol{b}\in\mathbf{R}^{C}$为偏置项; $\boldsymbol{x}$为输入的语义特征; $\boldsymbol{y}$为经过压缩后的输出特征。

压缩后的公共特征$\boldsymbol{W}'_{I\mathrm{c}}$被映射到一个共享的信道子集$\boldsymbol{H}_{\mathrm{c}}$, 复用相同的时频资源。映射函数可以定义为

$$\boldsymbol{y}_{\mathrm{c}} = \boldsymbol{H}_{\mathrm{c}}\boldsymbol{P}_{\mathrm{c}}\boldsymbol{W}'_{I\mathrm{c}} + \boldsymbol{n}_{\mathrm{c}} \tag{15}$$

式中: $\boldsymbol{H}_{\mathrm{c}}$为公共特征对应的信道矩阵; $\boldsymbol{P}_{\mathrm{c}}$为功率分配矩阵; $\boldsymbol{n}_{\mathrm{c}}$为共享信道中的高斯噪声, $\boldsymbol{n}_{\mathrm{c}}\sim N(0,\sigma^2)$; $\boldsymbol{y}_{\mathrm{c}}$为经过压缩后的公共输出特征。

压缩后的个性特征$\boldsymbol{W}'_{I\mathrm{i}}$被分配到独立的信道子集$H_{I\mathrm{i}}=\{\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots,\boldsymbol{H}_K\}$, K为个性特征的总个数, 每个个性特征通过独立的时频资源进行传输, 该过程表示为

$$\boldsymbol{y}_{k'} = \boldsymbol{H}_{k'}\boldsymbol{P}_{k'}\boldsymbol{W}'_{k'} + \boldsymbol{n}_{k'} \tag{16}$$

式中: $\boldsymbol{H}_{k'}$为第k'个个性特征的信道矩阵; $\boldsymbol{P}_{k'}$为独立的功率分配矩阵; $\boldsymbol{n}_{k'}$为对应信道的高斯噪声, $\boldsymbol{n}_{k'}\sim N(0,\sigma^2)$; $\boldsymbol{y}_{k'}$为经过压缩后第k'个个性输出特征。

根据上述分析, 不同的图像中存在大量公共特征, 因此, 在传输这些图像时, 所使用的带宽资源是可以重叠的。假设共有N'张图像需要进行传输, 不同图像$n'(1\leqslant n'\leqslant N')$在传输过程中占用的带宽资源表示为$b_{n'}$, 其对应的带宽开销为$B_{n'}=|b_{n'}|$。为进一步衡量传输过程中图像在频域的重复利用率, 定义语义信息重用率S_{or}对其进行度量[18], 表示为

$$S_{\mathrm{or}} = \frac{|b_1\cap b_2\cap\cdots\cap b_{N'}|}{|b_1|+|b_2|+\cdots+|b_{N'}|} \tag{17}$$

其反映了图像传输过程中共性特征的资源重用程度和带宽效率。

3　实验结果与分析

3.1　实验设置

为验证以 Swin Transformer 为骨干网络基于特征划分传输的无线图像传输系统的端到端传输性能, 对该系统与几个基准测试进行比较。

由于语义编、解码仅采用 2 个阶段的 Swin Transformer, 因此, 属于低计算度语义编、解码器, 在此选择低分辨率图像以降低计算负担并提高处理效率。本节的所有结果均使用 CIFAR10 数据集中尺寸为 32×32×3(高度×宽度×通道)的图像训练和评价的模型获得。训练和测试集由完全不同的图像组成, 分别包含 50 000 和 10 000 张图像。批处理大小为 128, 从中随机配对 64 对进行传输, 所有绘制结果都是从评估数据集上每个图像通道的 10 次实现中获得的平均值。

在训练阶段, 使用学习率为1×10^{-4}的 Adam 优化器。所有模型训练与部署均以 Pytorch 为架构实现, 模型使用单块 RTX 3090 GPU 训练。对于 CIFAR10 数据集, 在语义特征提取部分, 采用 2 个阶段, 并将窗口大小设置为 2。特别注意, 由于低分辨率图像本身包含较少的细节信息, 较小的窗口能够集中捕捉局部区域的关键特征, 因此, 选择较小的窗口有助于在低分辨率图像上高效提取局部特征[15]。

而端到端的图像传输包括信源信道编码和无线信道传输 2 个主要模块。因此, 对比方法包括: 基于 Swin Transformer 的无线图像传输方法 (记作 WITT)[16]; 基于 CNN 的 Deep JSCC 方法(记作 Deep JSCC)[11]; 经典分离式信源信道编码 BPG 方法(H.265 视频编码标准的帧内图像编码方法)[21], 并结合实际应用的 5G 标准低密度奇偶校验码 (low-density parity-check, LDPC) 信道编码[22](记为 BPG+LDPC)。在此, 采用块长度为 6 144 bit 的 5G LDPC 码, 支持不同的编码速率和正交幅度调制(quadrature amplitude modulation, QAM)。本文考虑了在 AWGN 信道和瑞利衰落信道下的性能测试。为确保实验设置的一致性和公平性, 信道输入的每 2 个连续实符号被合并为一个复信道输入符号, 并在信道中加入复高斯噪声。

本文研究的是非任务驱动的语义图像传输方

法，旨在优化图像传输质量和提高传输效率。实验使用广泛应用的像素级度量指标PSNR和结构相似性指数(structural similarity index measure, SSIM)对本文模型和其他端到端传输模型进行性能评估，并利用语义信息重用率对传输过程中图像在频域的重复利用率进行度量。

PSNR对应于像素级的L2欧氏距离，因此，当评估模型在PSNR上的表现时，将失真d设置为信源图像$\boldsymbol{I}$和重构图像$\hat{\boldsymbol{I}}$间的均方误差(mean squared error, MSE)：

$$R_{\mathrm{PSNR}}=10\lg\left(\frac{X_{\mathrm{MAX}}^2}{E_{\mathrm{MSE}}}\right) \tag{18}$$

式中：R_{PSNR}为峰值信噪比；E_{MSE}为均方误差；X_{MAX}为图像像素的最大可能值[11]，对于在单个颜色通道中具有8位像素的图像，X_{MAX}值为255。因此，最小化MSE导致最大化PSNR。

SSIM描述了接收端的重建图像与发送端的原始图像之间的相似性：

$$I_{\mathrm{SSIM}}(\boldsymbol{X},\hat{\boldsymbol{X}})=\frac{(2\mu_X\mu_{\hat{X}}+C_1)(2\sigma_{X\hat{X}}+C_2)}{(\mu_X^2+\mu_{\hat{X}}^2+C_1)(\sigma_X^2+\sigma_{\hat{X}}^2+C_2)} \tag{19}$$

式中：$\boldsymbol{X}$和$\hat{\boldsymbol{X}}$分别为原始图像和重建图像的局部区域；μ_X和σ_X^2分别为$\boldsymbol{X}$的均值和方差；$\sigma_{X\hat{X}}$为$\boldsymbol{X}$与$\hat{\boldsymbol{X}}$之间的协方差；C_1、C_2是为避免分母为0而引入的常数。

3.2 仿真结果分析

如图8所示，随着阈值增大，语义信息重用率逐渐上升。这表明，随着阈值的增大，更多的公共特征被提取，从而提高了带宽的重用效率。从图8可以看出，在较低阈值时，语义信息重用率提高近40%，说明阈值较低时，公共特征的提取数量较多，有效提升带宽利用率。而在较高阈值下，语义信息重用率的增长趋于平缓，仅提升约20%，这表明在较高阈值下，增加公共特征数量对带宽利用率的提升效果有限。因此，阈值的合理选择对于优化带宽利用和图像恢复质量至关重要，过高的阈值可能导致公共特征的过度重用，影响传输质量。

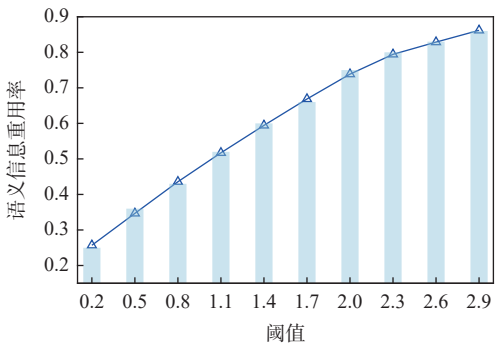


图8 语义信息重用率随阈值的变化

Fig. 8 Variation of semantic information reuse rate with threshold

图9展示了在AWGN信道下的图像传输质量。可知在不同的语义信息重用率下本文方法的图像传输质量。黑色线条表示带宽占用率，即在传输语义信息过程中所需要的实际带宽与原始带宽之间的比例。可以看出，随着语义信息重用率的增加，带宽占用率逐步降低。当传输相同的语义信息量时，带宽占用量仅为原始带宽占用量的一半。红色线条展示了本文方法在不同语义信息重用率下的性能变化情况。可以观察到，当语义信息重用率达到0.3时，图像传输质量接近于语义信息重用率为0时的图像传输质量，说明此时能够节省大约15%的带宽资源。然而，当语义信息重用率超过0.3时，尽管带宽节省更多，但恢复性能开始逐步下降。这表明重用语义信息在提升带宽效率的同时，也对图像恢复质量产生一定的影响。

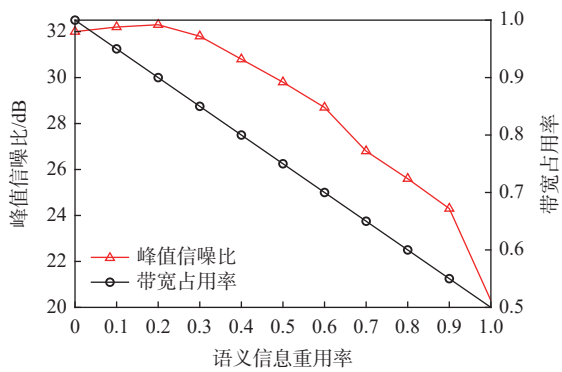


图9 图像传输质量随语义信息重用率的变化

Fig. 9 Variation of image transmission quality with semantic information reuse rate

图10展示了在AWGN信道下CBR=1/3时的图像传输质量。可知，在不同SNR条件下，不同的传输方法表现出了不同的图像传输质量。本文方法(语义信息重用率为0.3)与基于CNN的Deep JSCC方法相比，展现出了更好的性能表现，不仅节省了带宽，而且在不同SNR条件下都能够实现较

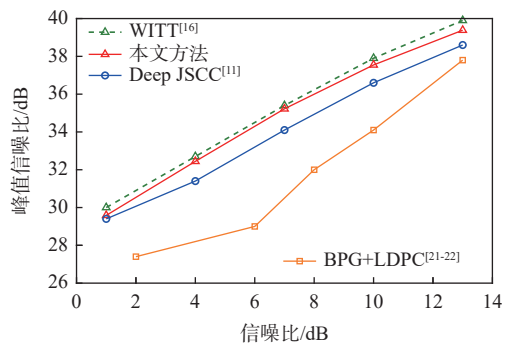


图10 AWGN信道下CBR=1/3时的图像传输质量

Fig. 10 Image transmission quality at CBR=1/3 over AWGN channel

高的 PSNR 值。此外, 本文方法表现性能较 WITT 方法有所下降, 其原因在于本文方法共性特征通过相同的时频资源进行传输, 可能导致信息丢失或传输不足, 特别是在信道质量较差时。相比之下, 不区分传输时, 共性和个性特征都能通过多次传输得到更多冗余, 增强了系统的鲁棒性, 避免了信息丢失。

图 11 展示了在 AWGN 信道下当 SNR=10 dB 时 PSNR 与 CBR 之间的关系。可以看出, 随着 CBR 值增加, 各方法的图像质量逐渐上升。此外, 本文方法(语义信息重用率为 0.3)在低 CBR 条件下, 相较于基于 CNN 的 Deep JSCC 方法及 BPG+LDPC 方法, 仍能保持较高的 PSNR 值, 表现出了较强的鲁棒性和适应性, 能够在低带宽环境下有效维持图像质量。尽管相较 WITT 方法, 本文方法在性能上有小幅度下降, 但本文方法通过在特征传输上进行有效区分, 减少了传输中的冗余信息, 从而节约了带宽资源。此外, 在 CBR = 0.5 时, Deep JSCC 方法和 BPG+LDPC 方法的 PSNR 值相交, 这是因为随着 CBR 的提高, Deep JSCC 方法的联合信源信道编码优势逐渐减弱, 而 BPG+LDPC 方法在较高压缩比时能更好地利用信道容量进行图像恢复。

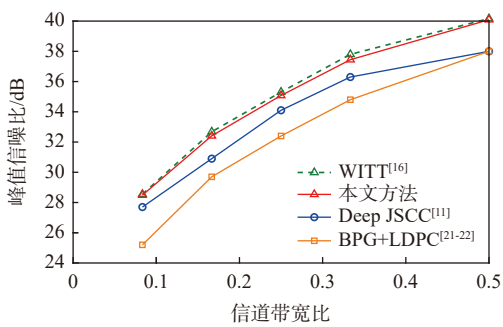


图 11　AWGN 信道下 SNR=10 dB 时的图像传输质量

Fig. 11　Image transmission quality at SNR=10 dB over AWGN channel

由图 12 可以看出, 在瑞利衰落信道下, 当 CBR=1/3 时, 不同 SNR 条件下的图像传输质量随 SNR 的提升而逐渐改善, 表现为 PSNR 值随着 SNR 的增加而上升, 即图像传输质量稳步提高。尽管本文方法(语义信息重用率为 0.3)相比 WITT 方法的传输质量略有下降, 但仍比 Deep JSCC 方法和 BPG+LDPC 方法表现出更佳的图像质量, 具有明显的相对优势。在较低 SNR 条件下, 由于噪声较大, 共性信息和个性信息可能会部分丢失或受到损坏, 进而对传输质量产生一定影响。不过总体而言, 本文方法在各类 SNR 环境下均能保持良好的图像质量, 从而提升了传输效果的稳定性。此外, 在 SNR = 11 dB 时, BPG+LDPC 方法与 Deep JSCC 方法的 PSNR 值接近, 这是由于 BPG+LDPC 方法在较高 SNR 下能更

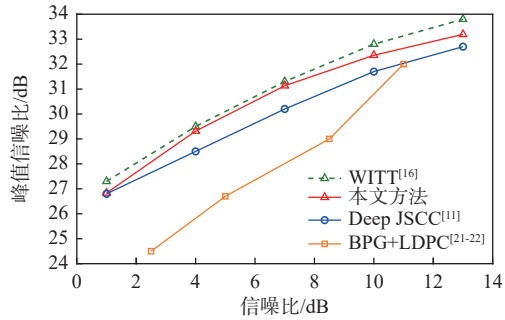


图 12　瑞利衰落信道下 CBR=1/3 时的图像传输质量

Fig. 12　Image transmission quality at CBR=1/3 over Rayleigh channel

好地发挥其信道编码的纠错能力, 导致其性能接近于 Deep JSCC 方法。

图 13 展示了在瑞利衰落信道下 SNR=3 dB 时 PSNR 与 CBR 的关系。可以看出, 本文方法(语义信息重用率为 0.3)对比 Deep JSCC 方法及 BPG+LDPC 方法, 都展示出了更好的性能表现。另外, 值得注意的是, 当 CBR 较大时, 本文方法对比 WITT 方法, 传输质量也非常接近, 这是由于在信道带宽较大的情况下, 系统的传输能力增强, 能够同时处理更多数据, 因此, 分开传输特征的方式对性能的影响变得较小。

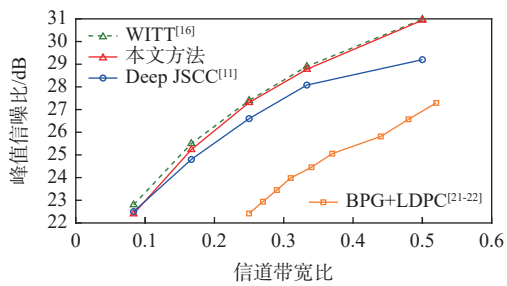


图 13　瑞利衰落信道下 SNR=3 dB 时的图像传输质量

Fig. 13　Image transmission quality at SNR=3 dB over Rayleigh channel

为更全面地评估本文方法的性能, 进一步通过 SSIM 对实验结果进行了分析。SSIM 作为一种衡量图像质量的感知度量, 能够较好地反映图像的结构相似性、亮度和对比度等感知特性。图 14 展示了在瑞利衰落信道下 CBR=1/3 时 SSIM 与 SNR 之间的关系。实验结果显示, 随着 SNR 的提升, 所有方法的 SSIM 值均呈现增长趋势, 本文方法(语义信息重用率为 0.3)在各 SNR 条件下的 SSIM 值始终优于 Deep JSCC 方法和 BPG+LDPC 方法。结合图 12 中的 PSNR 结果表明, 相比传统图像传输方法, Deep JSCC 方法在图像质量上表现更佳。这是因为传统方法优化 MSE, 而深度学习方法通过学习驱动, 更有效地保留和恢复图像细节, 提升了图像质量。

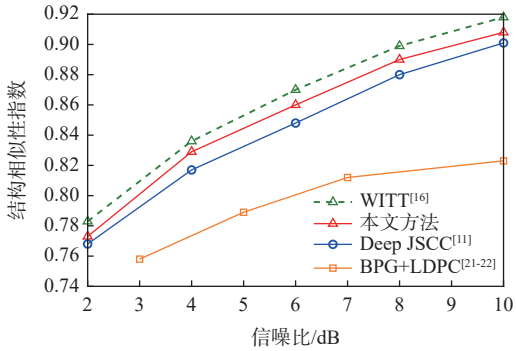


图 14 瑞利衰落信道下 CBR=1/3 时的 SSIM 性能

Fig. 14 SSIM performance at CBR=1/3 over Rayleigh channel

4 结 论

1）采用 Swin Transformer 作为网络架构，相比传统的基于 CNN 的 Deep JSCC 方法，本文方法在各 SNR 条件下都实现了更高的 PSNR。

2）在低 SNR 条件下，通过优化图像特征的传输方式，本文方法在保证图像质量的同时，节省了带宽。

3）通过区分公共特征和个性特征，本文方法有效降低了带宽需求，能够显著提升通信系统的整体性能，并降低传输成本。

因此，本文提出的基于 Swin Transformer 的共性和个性信息传输策略为图像传输系统的优化和性能提升提供了一种有效的解决方案。在后续研究中，可以针对基于 Swin Transformer 的多用户的图像传输问题进行探讨。

参考文献（References）

[1] Dai J C, Zhang P, Niu K, et al. Communication beyond transmitting bits: semantics-guided source and channel coding[J]. IEEE Wireless Communications, 2023, 30(4): 170-177.

[2] 张振国, 杨倩倩, 贺诗波. 基于深度学习的图像语义通信系统[J]. 中兴通讯技术, 2023, 29(2): 54-61.
Zhang Z G, Yang Q Q, He S B. Image semantic communication system based on deep learning[J]. ZTE Technology Journal, 2023, 29(2): 54-61(in Chinese).

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

[4] 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.

[5] 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(in Chinese).

[6] 张平, 戴金晟, 张育铭, 等. 面向语义通信的非线性变换编码[J]. 通信学报, 2023, 44(4): 1-14.
Zhang P, Dai J S, Zhang Y M, et al. Nonlinear transform coding for semantic communications[J]. Journal on Communications, 2023, 44(4): 1-14(in Chinese).

[7] Zhang Y C, Zhao H T, Wei J B, et al. Context-based semantic communication via dynamic programming[J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(3): 1453-1467.

[8] 江沸菠, 彭于波, 董莉. 面向 6G 的深度图像语义通信模型[J]. 通信学报, 2023, 44(3): 198-208.
Jiang F B, Peng Y B, Dong L. Deep image semantic communication model for 6G[J]. Journal on Communications, 2023, 44(3): 198-208(in Chinese).

[9] 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.

[10] Yoo H, Dai L L, Kim S, et al. On the role of ViT and CNN in semantic communications: analysis and prototype validation[J]. IEEE Access, 2023, 11: 71528-71541.

[11] Bourtsoulatze E, Burth Kurka D, Gündüz D. Deep joint source-channel coding for wireless image transmission[J]. IEEE Transactions on Cognitive Communications and Networking, 2019, 5(3): 567-579.

[12] Kurka D B, Gündüz D. DeepJSCC-f: deep joint source-channel coding of images with feedback[J]. IEEE Journal on Selected Areas in Information Theory, 2020, 1(1): 178-193.

[13] Xu J L, Tung T Y, Ai B, et al. Deep joint source-channel coding for semantic communications[J]. IEEE Communications Magazine, 2023, 61(11): 42-48.

[14] 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.

[15] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE Press, 2022: 9992-10002.

[16] Yang K, Wang S X, Dai J C, et al. WITT: a wireless image transmission transformer for semantic communications[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2023: 1-5.

[17] Nguyen L X, Tun Y L, Tun Y K, et al. Swin transformer-based dynamic semantic communication for multi-user with different computing capacity[J]. IEEE Transactions on Vehicular Technology, 2024, 73(6): 8957-8972.

[18] Chen G, Nan G S, Jiang Z H, et al. Lightweight and robust wireless semantic communications[J]. IEEE Communications Letters, 2024, 28(11): 2633-2637.

[19] Zhang P, Xu X D, Dong C, et al. Model division multiple access for semantic communications[J]. Frontiers of Information Technology & Electronic Engineering, 2023, 24(6): 801-812.

[20] Zhang P, Xu W J, Liu Y M, et al. Intellicise wireless networks from semantic communications: a survey, research issues, and challenges[J]. IEEE Communications Surveys & Tutorials, 2025, 27(3): 2051-2084.

[21] Albalawi U, Mohanty S P, Kougianos E. A hardware architecture for better portable graphics (BPG) compression encoder[C]//Proceedings of the 2015 IEEE International Symposium on Nanoelectronic and Information Systems. Piscataway: IEEE Press, 2015: 291-296.

[22] Richardson T, Kudekar S. Design of low-density parity check codes for 5G new radio[J]. IEEE Communications Magazine, 2018, 56(3): 28-34.

Swin Transformer-based semantic image transmission with model division multiplexing

PENG Yi[1, 2], SUN Chang[1, 2], YANG Qingqing[2, *], LI Hui[2], WANG Jianming[2]

(1. School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650031, China;
2. Yunnan Key Laboratory of Computer Technology Application, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Communication efficiency is greatly increased via semantic communication, a new paradigm for next-generation communication systems. However, previous works on semantic image transmission were mostly built upon convolutional neural networks (CNN). Traditional CNN-based models often struggle to comprehensively extract the semantic information embedded in images. To address this limitation, this paper adopts the Swin Transformer as the backbone network to effectively capture semantic information. Based on this, a novel semantic image transmission method leveraging model division is proposed. This method achieves the independent transmission of public and personalized features through feature separation. Specifically, after the input image undergoes joint source-channel coding (JSCC), the encoded semantic features are categorized into public features and personalized features, which are then transmitted separately. By enabling resource sharing for public features, the proposed method significantly reduces bandwidth requirements, thereby improving transmission efficiency. To evaluate the performance of the proposed model, experiments are conducted on the publicly available CIFAR10 dataset, comparing the proposed method with traditional approaches. The experimental results demonstrate that the proposed method effectively improves image transmission quality. Furthermore, it achieves superior performance in terms of transmission efficiency and bandwidth savings. In order to overcome bandwidth limitations in semantic picture transfer while preserving high-quality image reconstruction, this work offers a promising alternative.

Keywords: semantic communication; image transmission; model division multiple access; image reconstruction; Swin Transformer

Received: 2024-07-18; **Accepted**: 2024-12-20; **Published Online**: 2025-01-16 13:40

URL: link.cnki.net/urlid/11.2625.V.20250115.1641.001

Foundation items: National Natural Science Foundation of China (62461030); Key Program of the Yunnan Provincial Basic Research Plan (202401AS070105)

* **Corresponding author**. E-mail: 12309214@kust.edu.cn