doi:10.3969/j.issn.1003-3114.2025.05.014
引用格式:伍忠东,甘炳坤,王鹏波,等.基于多任务的图像语义传输方法[J].无线电通信技术,2025,51(5):1016-1024.
[WU Zhongdong, GAN Bingkun, WANG Pengbo, et al. Multi-task Based Approach for Semantic Transfer of Images[J]. Radio Communications Technology, 2025, 51(5): 1016-1024.]

基于多任务的图像语义传输方法

伍忠东,甘炳坤,王鹏波,苟敬聪,丁尚思
(兰州交通大学 电子与信息工程学院,甘肃 兰州 730070)

摘　要:近年来,基于 Transformer 的视觉模型,如 Swin Transformer,在视觉任务中展现出良好的前景,然而这些方法通常侧重于减少原始数据与重建数据之间的信号失真,而忽略感知质量。针对传统均方误差(Mean Square Error,MSE)损失难以反映图像感知与语义质量的不足,设计了 MSE 与学习感知图像块相似度(Learned Perceptual Image Patch Similarity,LPIPS)的加权组合损失函数,从而构建基于 Swin Transformer 的语义通信框架,称为融合感知损失的联合信源信道编码(Swin Transformer with LPIPS-based Joint Source-Channel Coding,STL-JSCC)方法,显著提升了图像重建质量与语义还原能力。在性能评估方面,设计了图像语义偏差值(Images Semantic Deviation,ISD)与语义相似度(Images Semantic Similarity,ISS)2 项指标,构建联合感知-语义评估体系,突破传统评价方法局限。实验结果表明,提出的 STL-JSCC 在各项指标上均优于其他模型,验证了所提方法在提升图像重建质量和语义提取能力上所具有的显著潜力和优势。

关键词:Swin Transformer;语义通信;学习感知图像块相似度;语义评估

中图分类号:TN914　　**文献标志码**:A　　**开放科学(资源服务)标识码(OSID)**:

文章编号:1003-3114(2025)05-1016-09

Multi-task Based Approach for Semantic Transfer of Images

WU Zhongdong, GAN Bingkun, WANG Pengbo, GOU Jingcong, DING Shangsi
(School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China)

Abstract: In recent years, Transformer-based visual models (e.g., Swin Transformer) show good prospects in visual tasks, however, these methods usually focus on reducing signal distortion between original and reconstructed data, while ignoring perceptual quality. Considering that the conventional Mean Square Error (MSE) loss fails to reflect perceptual and semantic quality effectively, we propose a weighted loss function combining MSE and Learned Perceptual Image Patch Similarity (LPIPS), and accordingly construct a Swin Transformer-based semantic communication framework, called Swin Transformer with LPIPS-based Joint Source-Channel Coding (STL-JSCC) method, which significantly enhances image reconstruction quality and semantic consistency. For performance evaluation, two semantic-aware metrics are introduced: the Images Semantic Deviation (ISD) value and Iamges Semantic Similarity(ISS). These indicators form a joint perceptual-semantic evaluation system, which breaks through the limitations of traditional evaluation methods. Experimental results show that the proposed STL-JSCC outperforms other models in all the indexes, verifying the significant potential and advantages of the proposed method in improving the image reconstruction quality and semantic extraction capability.

Keywords: Swin Transformer; semantic communication; LPIPS; semantic evaluation

0　引言

近年来,人工智能和通信技术的融合催生了一种新的通信范式——语义通信。作为一种基于深度学习的端到端传输系统[1],语义通信是一种新型通信范式,旨在传输信息时不仅保留数据的原始内容,更注重其语义信息的有效传递[2]。与传统通信系统(基于 Shannon 分离定理[3],分离源编码和信道编码的方法)不同,语义通信通过联合信源信道编码(Joint Source-Channel Coding,JSCC)实现端到端优化,强调在复杂信道条件下保持信息的语义一致性,从而提升传输效率和重建质量[4]。

收稿日期:2025-05-06
基金项目:甘肃省科技重大专项(22ZD6GA041)
Foundation Item:Gansu Provincial Major Science and Technology Special Project (22ZD6GA041)

在图像传输任务中,近年来一种基于深度学习的端到端优化的 JSCC 方法——DeepJSCC,成为语义通信领域研究的热点方向。该方法通过神经网络对图像编码和信道映射过程进行联合学习,在保证语义保真度的同时显著增强了通信系统在复杂环境下的鲁棒性[5-11]。例如,Bourtsoulate 等[5]开创性地使用 DeepJSCC 实现端到端图像传输,后续研究在信道自适应[12]、信噪比联合优化[13]等方面持续改进。针对卷积神经网络(Convolutional Neural Network,CNN)局部感受野的局限,Vision Transforme[14]在视觉任务中展现出独特优势:Yoo 等[15]首次将 ViT 的架构引入无线图像传输(SemViT),在语义通信领域取得较为出色的成果。后续对 ViT 网络进行改良,使得在图像处理任务中有很大的提升[16]。随之,Yang 等[17]采用 Swin Transformer 构建无线图像传输转换器(Wireless Image Transmission Transformer,WITT)。相比传统卷积网络,Transformer 架构能有效捕捉全局语义特征,显著提升复杂场景的传输性能。

虽然上述方法取得了显著进展,可以保证较好的语义通信效果,但是优化目标仍存在很多局限。

① 编解码器。基于 Swin Transformer 的编码器虽然具备良好的局部特征提取能力,但在建模图像全局依赖方面存在一定局限性[18]。

② 模型训练。现有工作普遍以 MSE、峰值信噪比(Peak Signal to Noise Ratio,PSNR)、结构相似性指数(Structural Similarity Index,SSIM)等传统失真度量作为优化目标,但这些指标仅衡量像素级保真度,与人类视觉系统的感知质量(Perceptual Quality)相关性较弱[19]。过度优化 MSE 甚至会导致解码图像出现语义信息失真[20]。

③ 评价指标。当前语义通信系统常依赖传统评估指标或特定任务指标,前者无法反映语义保真度,后者则因任务差异性难以通用化,以至于评估体系不完善。

针对以上问题,本文构建了用于语义通信的感知-失真联合优化的通信架构:在信号表征层面,设计了 MSE 与 LPIPS[21]的加权组合损失函数,通过像素精度与感知相似度的自适应加权机制,实现低阶纹理特征与高阶语义特征的协同优化,有效提升了图像语义重建效果;在系统评估层面,提出联合感知-语义的多任务评估体系,以 ISD 量化语义信息传递的保真度,以 ISS 评估任务相关的语义一致性。结合基于传统指标的质量评估,一定程度上统一了语义通信系统"保真重建""任务适配"的评估范式,可更全面、准确地度量图像语义通信系统在不同任务中的表现。

1 基础理论

1.1 系统模型

本文利用联合源信道编码方法,构建的端到端图像语义通信系统,包含编码器、信道、解码器 3 个核心模块。传统方法通常采用"先源编码、后信道编码"的分离式架构,如图 1 所示。发送端首先对输入图像进行特征提取和语义编码,然后经过独立的信道编码模块进行调制处理,最终将编码后的信息传输至信道;接收端则依次完成信道解码和语义解码;接收端则经信道解码器逆处理后再由语义解码器根据接收到的语义信息进行图像重建。如图 2 所示,本文使用的方法是编码器、解码器模块将语义编码和信道编码过程结合,以实现语义编码与信道编码高效协同。

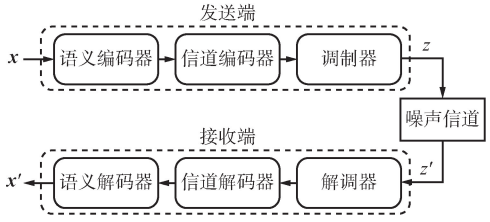

图 1 传统处理流程

Fig. 1 Taditional processing flow

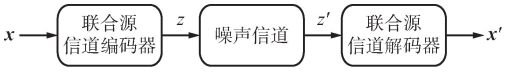

图 2 JSCC

Fig. 2 JSCC

用于无线图像传输的 STR-JSCC 网络架构中实现无线图像传输的语义-信道联合编码器如图 3 所示,编码器采用多级特征压缩:先通过补丁嵌入(Patch Embedding,PE)模块将输入的 RGB 图像 $\boldsymbol{x}\in R^{H\times W\times 3}$ 划分为不重叠的补丁 $P_1\in(H/2)\times(W/2)$。再通过补丁合并(Patch Merging,PM)模块拼接和线性降采样,以及级联 Swin Transformer 模块进行特征变换,以保持特征分辨率。编码过程分多个阶段,每个阶段集成 PM 模块和对应数量的 Swin Transformer 模块、嵌入维度和注意力头数。随着阶段推进,图像特征分辨率呈指数级下降,不同分辨率图像所需阶段数量亦有所不同。一般来说,更高分辨率的图像需要更多的阶段数量。编码过程如下:

$$y=E_\theta(\boldsymbol{x},SNR,R)\ , \tag{1}$$

式中:R 为传输图片集,E_θ 为可训练的编码器,输入图像由向量 $\boldsymbol{x} \in R^{H\times W\times 3}$ 表示,SNR 为通道信噪比,$R=K/(H\times W\times 3)$ 表示带宽压缩比。

编码器输出的语义特征输入到通道 ModNet 中,经过全连接层处理,与信噪比对应的 m 维特征向量连续融合。经非线性激活函数 Sigmoid 函数处理后,将特征与输入特征的残差相结合,从而得到调制后的输出。调制信号在经过功率归一化处理后被分解为 I/Q 分量,并通过无线信道进行传输,信号在此过程中受到噪声干扰。

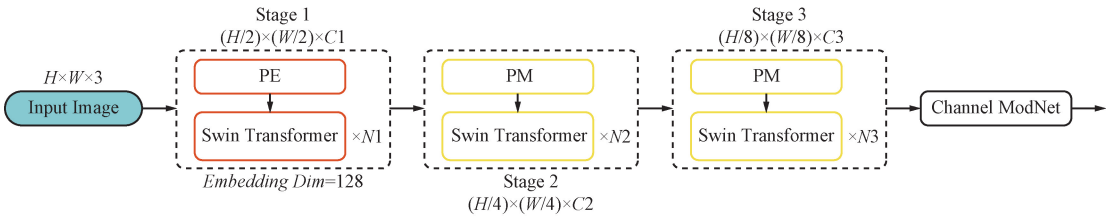

图 3　编码器结构

Fig. 3　Encoder structure

本文主要关注 2 类典型的信道模型:加性高斯白噪声(Additive White Gaussian Noise,AWGN)信道与瑞利衰落信道。信号在含噪信道中的传输过程可以建模为:

$$\tilde{y} = h \cdot y + n , \tag{2}$$

式中:h 表示服从瑞利分布的信道增益,n 为 AWGN。对于复数信道而言,n 的实部和虚部均独立服从正态分布。

解码器 D_ψ 与编码器 E_θ 具有对称结构,如图 4 所示。通过补丁划分(Patch Reverse Merging, PRM)模块的线性上采样与逆向特征变换重建图像。

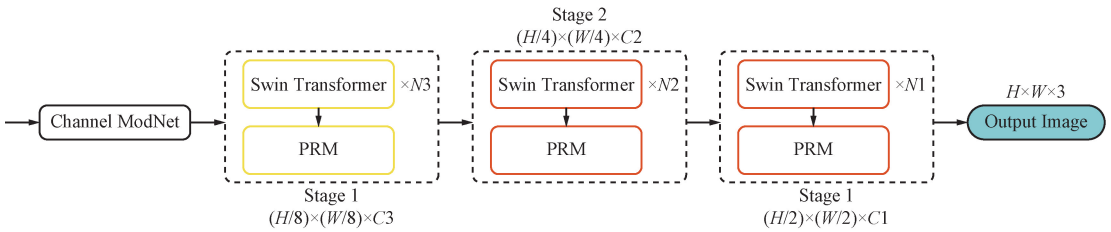

图 4　解码器结构

Fig. 4　Decodder structure

输出的重建图像表示为:

$$\tilde{\boldsymbol{x}} = D_\psi(\tilde{y}, SNR, R) 。 \tag{3}$$

训练优化以端到端 MSE 最小化为目标,联合约束编码器与解码器的网络参数 θ 和 ψ,平衡压缩效率与重建质量:

$$\underset{\theta,\psi}{\operatorname{argmin}}\ MSE(\boldsymbol{x}, \tilde{\boldsymbol{x}}) , \tag{4}$$

式中:$MSE(\ )$ 表示原始图像和重建图像的失真测量。

传统的图像语义通信模型通常仅依赖 MSE 来更新整个网络模型,但这种方法存在一个明显的问题,即在上采样过程中,图像的细节冗余丢失,从而导致重建图像的语义信息出现失真。为了尽可能地重建出与原图像相似的图像,采用学习到的 LPIPS 损失[21]与 MSE 组合成的加权损失,作为图像语义通信模型的目标函数,即:

$$loss = \min_{\theta,\psi}\ \gamma_1 MSE(\boldsymbol{x}, \tilde{\boldsymbol{x}}) + \gamma_2 L(\boldsymbol{x}, \tilde{\boldsymbol{x}}) , \tag{5}$$

式中:γ_1 和 γ_2 为 MSE 和 LPIPS 损失的权重。这种加权损失策略能够更有效地捕捉图像的语义特征,从而提升重建图像的整体质量与语义一致性。

1.2　Swin Transformer 模块

如图 5 所示,Swin Transformer 模块采用了窗口多头自注意力(Windowed Multi-head Self-Attention, W-MSA)和移位窗口多头注意力(Shifted Window Multi-head Self-Attention, SW-MSA),以交替的方式成对使用[22]。具体而言,W-MSA 将图像划分为多个不相交的局部窗口,并在每个窗口内执行自注意力计算,有效减少计算复杂度。而 SW-MSA 则通过窗口偏移操作实现跨窗口的信息交互,弥补 W-MSA 仅在局部窗口内计算的限制,使模型能够建模更长距离的全局依赖关系。

每个 Swin Transformer 模块内部还包含残差连接和前馈神经网络多层感知器(Multilayer Perceptron, MLP)模块,用于提升模型和特征表达能力。对于

2 个连续 Swin Transformer 模块的计算如下所示:

$$\tilde{y}^k = W-MSA(LN(\tilde{y}^{k-1})) + y^{k-1}, \quad (6)$$

$$y^k = MLP(LN(\tilde{y}^k)) + \tilde{y}^k, \quad (7)$$

$$\tilde{y}^{k+1} = SW-MSA(LN(y^k)) + y^k, \quad (8)$$

$$y^{k+1} = MLP(LN(\tilde{y}^{k+1})) + \tilde{y}^{k+1}, \quad (9)$$

式中:$\tilde{y}^k$ 和 y^k 分别表示 W-MSA 模块、SW-MSA 和 MLP 模块的输出特征。LN(Layer Normalization)表示层归一化。

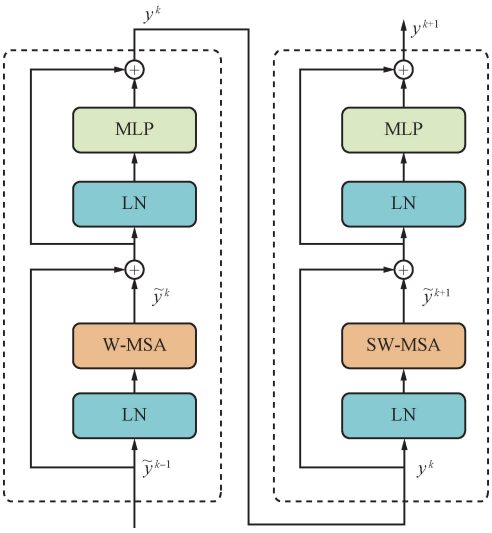

图 5 Swin Transformer 模块

Fig. 5 Swin Transformer block

2 评估方法与实验验证

2.1 评估方法

在 JSCC 算法和其他基准方案中,均采用 PSNR 和 MS-SSIM 对图像重建质量进行客观评价[23],在此基础上引入 ISD 和 ISS 指标对原图像与重建图像进行语义层级上的评估,从而构建完善的评估体系,对模型的通信效果进行全面评估。

PSNR 通过计算原始图像与重建图像之间的均方误差来量化图像之间的差异,PSNR 的值越高,重建图像的质量越高:

$$PSNR = 10\lg\left(\frac{255^2}{MSE}\right)。 \quad (10)$$

MS-SSIM 是一种多尺度结构相似性方法,通过对多个尺度的图像进行比较,从而获取全面的结构信息评估图像质量[24]。相比于 PSNR,MS-SSIM 考虑了图像的结构信息,能够更好地反映人眼对图像质量的感知,其数值越大表示图像的结构相似度越高,即对图像的感知质量越高。为了更直观地观察比较,将其转换为分贝(dB)的形式:

$$MS\text{-}SSIM(\text{dB}) = -10\lg(1-MS\text{-}SSIM)。 \quad (11)$$

V_{ISD} 旨在衡量重建图像相较于原图像分类结果的准确性。利用图像分类器对原图像和重建图像进行识别分类,并根据识别结果系统性地计算准确率。假设数据集有 N 个类别,计算如下:

$$V_{ICA} = \frac{1}{N}\sum_i^N 1(P_i(\boldsymbol{x}') = P_i(\boldsymbol{x})), \quad (12)$$

$$V_{ISD} = \lg(1 - V_{ICA}), \quad (13)$$

式中:$P_i(\boldsymbol{x})$ 和 $P_i(\boldsymbol{x}')$ 分别为原图像 $\boldsymbol{x}$ 的真实类别和重建图像 $\boldsymbol{x}'$的预测类别。V_{ICA} 为重建图像的相对分类准确率,当 V_{ISD} 越小,分类准确率越高,说明通过该方法传输语义信息的准确率越高;反之则说明其不能准确地传输语义信息。

V_{ISS} 旨在利用特征提取任务的特性,通过语义特征角度来评估重建图像与原始图像的语义一致性。具体方法是对原始图像和重建图像进行特征提取,并计算二者特征向量的余弦距离,从而客观衡量它们在语义空间中的相似程度:

$$V_{ISS} = \cos(f(\boldsymbol{x}), f(\boldsymbol{x}')), \quad (14)$$

式中:cos( )表示余弦距离计算,$f(\boldsymbol{x})$ 和 $f(\boldsymbol{x}')$ 分别表示对原图像 $\boldsymbol{x}$ 和重建图像 $\boldsymbol{x}'$进行特征提取得到的特征向量。当 V_{ISS} 值较高时,表明二者的特征相似度高,即语义更加接近;反之,则意味着重建图像无法正确还原原图像的语义特征。

2.2 实验设置

实验环境:基于 PyTorch 2.0.1 + CUDA 11.8,使用 Python 3.10。

数据集:使用 2 个公开图像数据集进行训练与评估。第一个为公开的 CIFAR-10 图像数据集,每张图片为 3×32 pixel×32 pixel 的 RGB 图片,包含了 10 个不同类别,每个类别有 6 000 张图像,其中 5 000 张用于训练,1 000 张用于测试。训练时,数据会被翻转,以扩展 10 倍,避免过拟合。

为验证模型在高分辨率图像上的表现,本文引入了 BIRDS-400 数据集。该数据集包含 400 个鸟类类别,图像分辨率为 224 pixel×224 pixel,类别细粒度划分明显。训练与测试样本按照标准划分使用,用于评估模型在复杂语义和高分辨率场景下的性能。

实验采用 DeepJSCC[5]、WITT[17]作为对照的基准模型,使用 Adam 优化器[25],学习率设置为 0.000 1,数据批次大小为 128,信噪比为 0~25 dB,5 步长分布。

2.3 实验验证

2.3.1 通信质量评估

实验采用 PSNR、MS-SSIM 作为评估指标,以评

估实验所用模型的通信质量,实验结果如图 6~图 9 所示。

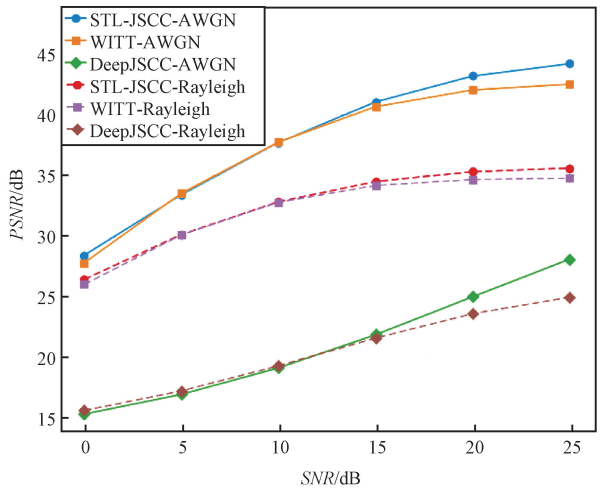

图 6　CIFAR10 上不同信道下的 PSNR 与 SNR 关系

Fig. 6　Plot of PSNR vs. SNR for different channels on CIFAR10

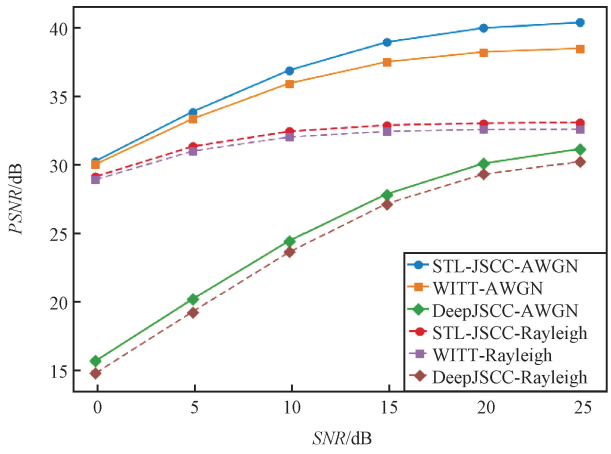

图 7　BIRDS-400 上不同信道下的 PSNR 与 SNR 关系

Fig. 7　Plot of PSNR vs. SNR for different channels on BIRDS-400

由图 6 和图 7 可以看出,随 SNR 增加,各模型的 PSNR 值均呈上升趋势,说明更高的 SNR 有助于提升图像重建质量;相比于其他模型,STL-JSCC 在全 SNR 范围内始终保持最高的 PSNR 值,特别是在高 SNR 条件下,其性能优势更为明显,表明其具有更高的通信质量。此外,在 AWGN 信道下,STL-JSCC 的性能显著优于 Rayleigh 信道,其 PSNR 上升速度更快,表明其具有更强的鲁棒性。

如图 8 和图 9 所示,STL-JSCC 在高 SNR 条件下的 MS-SSIM 值明显高于 WITT,表明其在细节结构还原方面的能力更强;而在低 SNR 环境下,二者性能相近,说明在恶劣信道下二者均具备一定鲁棒性。特别是在高 SNR 下的 AWGN 信道中,STL-JSCC 的 MS-SSIM 达到较高的水平,表明其在稳定信道条件下能够实现较高质量的图像重建。

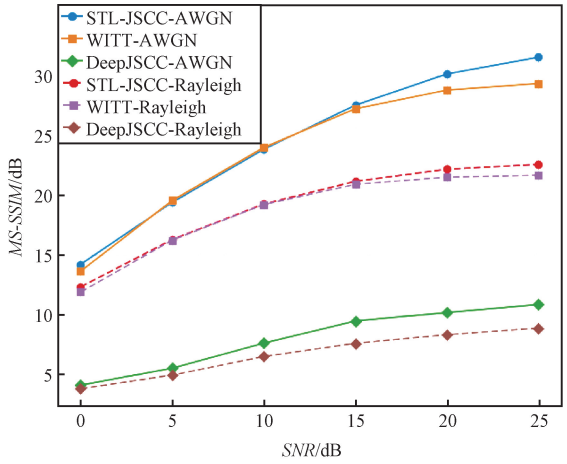

图 8　CIFAR10 上不同信道下的 MS-SSIM 与 SNR 关系

Fig. 8　Plot of MS-SSIM vs. SNR for different channels on CIFAR10

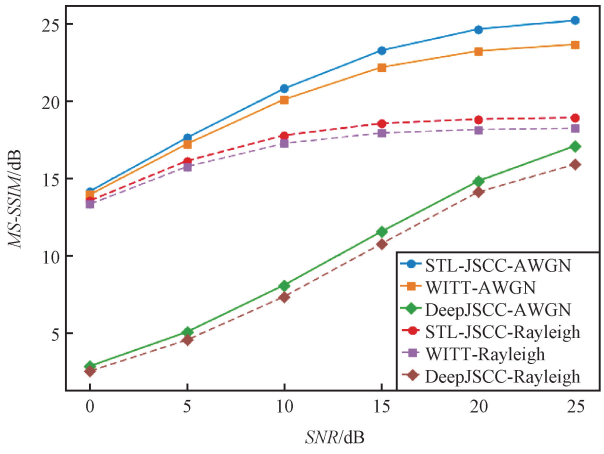

图 9　BIRDS-400 上不同信道下的 MS-SSIM 与 SNR 关系

Fig. 9　Plot of MS-SSIM vs. SNR for different channels on BIRDS-400

2.3.2　语义评估

为了评估重建图像与原图之间的语义一致性,本文采用迁移学习策略,使用具有预训练权重的深度神经网络 ResNet50[26],训练得到基于 CIFAR10 数据集预训练的深度神经网络作为公共图像分类器;冻结其特征层作为特征提取器以实现有效的特征提取,以此作为语义一致性评价的基础。

如表 1、表 2 所示,在不同数据集上,*SNR* = 25 dB 时,本文方法在 AWGN 和 Rayleigh 信道下的

图像重建质量和语义一致性指标均优于对比模型。

具体而言,在 CIFAR10 数据集上,AWGN 信道下的实验结果显示 $PSNR$ = 44.17 dB、$MS\text{-}SSIM$ = 31.56 dB,V_{ISD} = −1.26、V_{ISS} = 2.96,表明图像重建质量的提升在语义层面表现为更低的 ISD 和更高的语义相似度。

在高分辨率的 BIRDS-400 数据集上,尽管整体提升幅度相对较小,如 $PSNR$ = 40.33 dB、$MS\text{-}SSIM$ = 25.16 dB,V_{ISD} = −1.18,V_{ISS} = 2.29,仍可看出 ISD 降低、ISS 上升的趋势,与 CIFAR-10 数据集上的实验结果保持一致,说明所提出的语义评估指标在不同分辨率和图像复杂度下依然具备良好的一致性与稳定性,能够对模型实现更全面的评估。

表 1　在 CIFAR10 上不同通信模型的实验结果

Tab. 1　Experimental results for different communication models on CIFAR10

模型	信道	$PSNR$/dB	$MS\text{-}SSIM$/dB	V_{ISD}	V_{ISS}
DeepJSCC	AWGN	28.13	10.88	−0.37	0.77
	Rayleigh	24.99	8.92	−0.30	0.64
WITT	AWGN	42.49	29.37	−1.02	2.56
	Rayleigh	34.76	21.70	−0.68	1.74
STL-JSCC	AWGN	**44.17**	**31.56**	**−1.26**	**2.96**
	Rayleigh	**35.52**	**22.60**	**−0.74**	**1.88**

表 2　在 BIRDS-400 上不同通信模型的实验结果

Tab. 2　Experimental results for different communication models on BIRDS-400

模型	信道	$PSNR$/dB	$MS\text{-}SSIM$/dB	V_{ISD}	V_{ISS}
DeepJSCC	AWGN	31.10	17.05	−0.86	1.30
	Rayleigh	30.19	15.86	−0.82	1.24
WITT	AWGN	38.45	23.60	−1.15	2.20
	Rayleigh	32.56	18.19	−0.89	1.54
STL-JSCC	AWGN	**40.33**	**25.16**	**−1.18**	**2.29**
	Rayleigh	**33.04**	**18.86**	**−0.91**	**1.56**

为了验证模型的语义信息传输准确率,通过图像分类任务衡量语义传输的一致性。不同的通信模型在不同信道条件下的性能表现如图 10 和图 11 所示。可以看出,随着 SNR 增加,各模型的 ISD 整体呈下降趋势。相比于 WITT,STL-JSCC 在不同 SNR 条件下均展现出更优的语义保持能力,且显著优于 DeepJSCC。

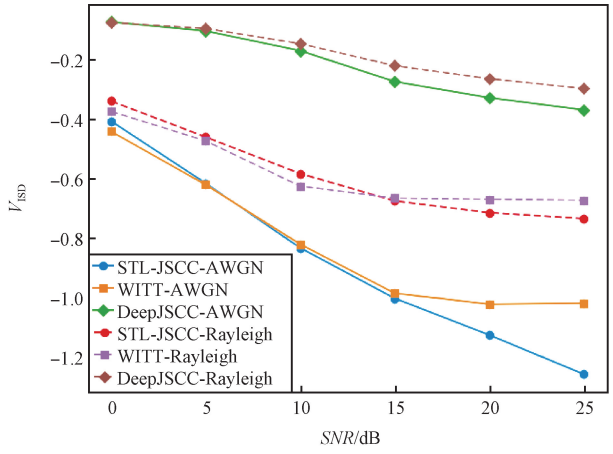

图 10　在 CIFAR10 上不同信道下的 V_{ISD} 与 SNR 关系

Fig. 10　Plot of V_{ISD} vs. SNR for different channels on CIFAR10

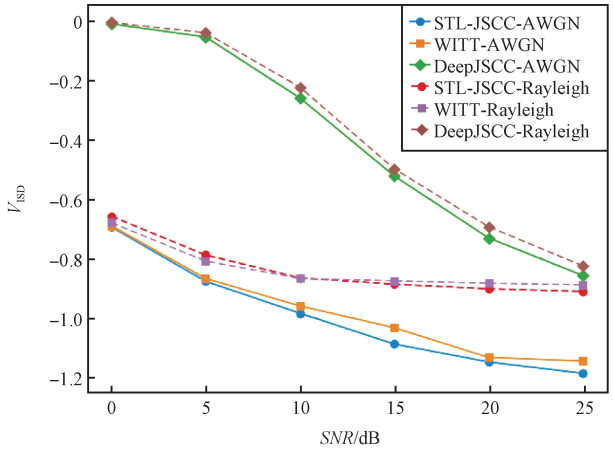

图 11　在 BIRDS-400 上不同信道下的 V_{ISD} 与 SNR 关系

Fig. 11　Plot of V_{ISD} vs. SNR for different channels on BIRDS-400

无论是在 AWGN 信道还是在 Rayleigh 信道上,STL-JSCC 在图像分类任务上的语义表现均优于其他基准模型,其 ISD 始终保持较低水平,说明该模型在图像重建过程中更好地保留了原图的语义信息。这一结果充分验证了 STL-JSCC 在保持图像语义信息方面的优越性和稳定性。

为了进一步验证模型能够更好地还原图像的语义,通过特征提取任务对原图像和重建图像的语义特征进行比较,如图 12 和图 13 所示。对比了不同模型在不同信道、SNR 条件下的特征相似度的变化,为了方便比较,计算余弦距离的对数并取正值作为实验指标。可以看出,随着 SNR 增加,各模型语义相似度整体呈上升趋势,其中 STL-JSCC 的语义相似度提升最为显著,明显优于其他基准模型,而后者的提升幅度相对较小。在 $SNR \geqslant 15$ dB 时,STL-JSCC 在不同信道中均取得最高的语义相似度,表现最为优异。值得注意的是,在 Rayleigh 信道下 STL-JSCC 与 WITT 模型之间的性能差距有所缩小,但整体语义相似度明显低于 AWGN 信道,反映出信道衰落对语义传输的影响。各模型在不同条件下的表现趋势均较为稳定,进一步验证了 STL-JSCC 在复杂条件下的鲁棒性和语义保持能力。

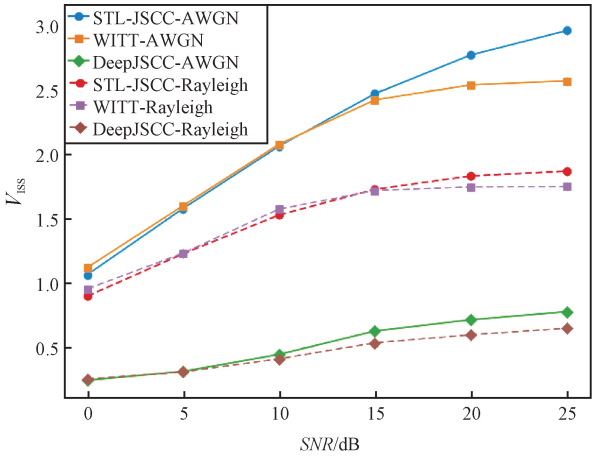

图 12　在 CIFAR10 不同信道下的 V_{ISS} 与 SNR 关系

Fig. 12　Plot of V_{ISS} vs. SNR for different channels on CIFAR10

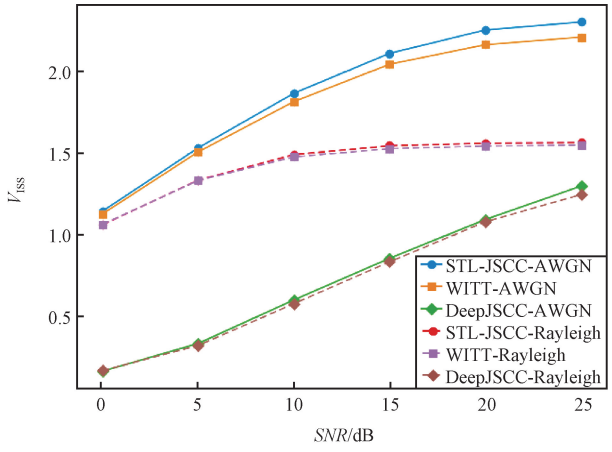

图 13　在 BIRDS-400 上不同信道下的 V_{ISS} 与 SNR 关系

Fig. 13　Plot of V_{ISS} vs. SNR for different channels on BIRDS-400

2.3.3　消融实验

为验证本文所引入加权损失函数的必要性及语义评估指标的有效性与合理性,在 $SNR = 25$ dB 和 AWGN 信道条件下,开展了不同权重分配的消融实验,结果如表 3 所示。

表 3　消融实验结果

Tab. 3　Results of ablation experiments

数据集	*MSE*	*LPIPS*	*PSNR*/dB	*MS-SSIM*/dB	V_{ISD}	V_{ISS}
CIFAR10	1.0	0	42.49	29.37	−1.02	2.56
	0.3	0.7	35.01	20.31	−0.62	1.43
	0.5	0.5	**44.17**	**31.56**	**−1.26**	**2.96**
BIRDS-400	1.0	0	38.45	23.60	−1.15	2.20
	0.3	0.7	37.80	22.69	−1.08	1.82
	0.5	0.5	**40.33**	**25.16**	**−1.18**	**2.29**

实验结果表明,传统基于 MSE 的训练目标侧重于像素级误差的最小,难以保持语义一致性。引入感知损失后,尽管 PSNR 提升有限,但 V_{ISS} 和 MS-SSIM 显著改善,说明该损失组合有助于提升语义一致性与人眼感知质量。合理分配 MSE 与感知损失的权重是提升语义通信系统性能的关键,纯 MSE 损失虽能优化传统指标,但语义保持能力不足;而过度依赖感知损失会导致重建质量下降。通过均衡权重

设计,STL-JSCC 模型在 PSNR、MS-SSIM 传统指标与 V_{ISD}、V_{ISS} 语义指标上均取得最优表现,证实了加权损失函数的合理性和必要性。

以上消融实验充分证明,本文提出的感知损失设计在提升语义通信模型的图像还原质量和语义保真度方面具有显著优势。此外,LPIPS 损失在低分辨率(32 pixel×32 pixel)场景中对语义和感知质量的提升更为显著,在高分辨率(224 pixel×224 pixel)场景中,其边际增益相对较小。这可能是由于高分辨率图像中结构细节更多,模型已具备更强的重建能力,感知损失的优势相对弱化。

2.3.4 模型复杂度与实时性分析

为全面评估所提出的基于 Swin Transformer 的 STJSCC 通信模型在不同分辨率场景下的计算复杂度与实时性表现,本文分别对低分辨率(32 pixel×32 pixel)和高分辨率(224 pixel×224 pixel)输入图像进行了建模与统计分析。由于模型在不同输入尺度下采用了分辨率自适应结构,编码器与解码器的实际深度与通道数均有所调整,因而在复杂度与延迟方面也表现出显著差异,结果如表 4 所示。在低分辨率(32 pixel×32 pixel)输入条件下,模型的总的每秒浮点运算数(Floating Point Operations Per Second,FLOPs)约为 0.77 G,参数量约为 13.75 M,平均单张图像推理时间为 12.60 ms。而在高分辨率(224 pixel×224 pixel)输入条件下,随着模型结构进一步加深,FLOPs 提升至 69.71 G,参数量为 71.66 M,平均推理时间为 28.31 ms。

上述结果表明,STJSCC 模型在不同输入分辨率下均具有良好的计算效率与实时性能。

表 4 复杂度与实时性分析结果

Tab. 4 Analysis results of complexity and real-time performance

图像尺寸/pixel	深度	*FLOPs*/G	参数量/M	推理时间/ms
32×32	2	0.77	13.75	12.60
224×224	3	69.71	71.66	28.31

3 结束语

针对传统 Transformer 模块依赖的传统 MSE 损失仅关注像素层级的误差,难以感知语义信息,本文设计了感知-失真联合优化的通信架构 STL-JSCC,构建基于 MSE 与 LPIPS 的加权组合损失函数,兼顾低级像素一致性与高级语义结构一致性,显著提升了图像重建的真实感与语义保真度。为了更全面地衡量图像语义传输质量,提出了结合 ISD 与 ISS 的评估方法,构建联合感知-语义的评估体系,从而突破了传统评估指标仅关注视觉质量的限制。实验结果表明,在 AWGN 与 Rayleigh 信道中,STL-JSCC 在 PSNR、MS-SSIM、V_{ISD}、V_{ISS} 等指标上全面优于现有方案,验证了所提方法在图像语义保真和传输鲁棒性方面的优势。

未来研究尝试提升低信噪比条件下的传输质量,还将探索有效的模型压缩技术,通过降低发送端和接收端的计算开销,在提升通信性能的同时缩减时延,确保语义通信系统的实时性。

参 考 文 献

[1] YE H,LIANG L,LI G Y,et al. Deep Learning-based End-to-End Wireless Communication Systems with Conditional GANs as Unknown Channels[J]. IEEE Transactions on Wireless Communications,2020,19(5):3133-3143.

[2] 陈建侨,马楠,许晓东,等. 面向语义通信的信道知识库构建与信道处理研究综述[J]. 无线电通信技术,2024,50(3):519-527.

[3] SHANNON C E. A Mathematical Theory of Communication[J]. The Bell System Technical Journal,1948,27(3):379-423.

[4] FRESIA M,PEREZ-CRUZ F,POOR H V,et al. Joint Source and Channel Coding[J]. IEEE Signal Processing Magazine,2010,27(6):104-113.

[5] BOURTSOULATE E,KURKA D B,GUNDUZ D. Deep Joint Source-channel Coding for Wireless Image Transmission[J]. IEEE Transactions on Cognitive Communications and Networking,2019,5(3):567-579.

[6] TUNG T Y,KURKA D B,JANKOWSKI M,et al. DeepJSCC-Q: Constellation Constrained Deep Joint Source-Channel Coding[J]. IEEE Journal on Selected Areas in Information Theory,2022,3(4):720-731.

[7] TUNG T Y,GUNDUZ D. DeepWiVe: Deep-learning-aided Wireless Video Transmission[J]. IEEE Journal on Selected Areas in Communications,2022,40(9):2570-2583.

[8] YANG M Y,BIAN C H,KIM H S. OFDM-guided Deep Joint Source Channel Coding for Wireless Multipath Fading Channels[J]. IEEE Transactions on Cognitive Communications and Networking,2022,8(2):584-599.

[9] KURKA D B,GUNDUZ D. Bandwidth-agile Image Transmission with Deep Joint Source-channel Coding[J]. IEEE Transactions on Wireless Communications,2021,20(12):8081-8095.

[10] XU J L, AI B, C 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.

[11] 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.

[12] YANG M Y, KIM H S. Deep Joint Source-channel Coding for Wireless Image Transmission with Adaptive Rate Control[EB/OL]. (2021-10-09)[2025-04-29]. https://arxiv.org/abs/2110.04456.

[13] ZHANG W Y, ZHANG H J, MA H, et al. Predictive and Adaptive Deep Coding for Wireless Image Transmission in Semantic Communication[J]. IEEE Transactions on Wireless Communications, 2023, 22(8): 5486-5501.

[14] HAN K, WANG Y H, CHEN H T, et al. A Survey on Vision Transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.

[15] YOO H J, DAI L L, KIM S K, et al. On the Role of ViT and CNN in Semantic Communications: Analysis and Prototype Validation[J]. IEEE Access, 2023, 11: 71528-71541.

[16] 刘铁,段勇. 融合 CNN 和 Transformer 的机器人室内场景识别[J]. 电子测量与仪器学报,2023,37(5): 223-229.

[17] YANG K, WANG S X, DAI J C, et al. WITT: A Wireless Image Transmission Transformer for Semantic Communications[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing. Rhodes Island: ICASSP, 2023: 1-5.

[18] 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.

[19] BLAU Y, MICHAELI T. Rethinking Lossy Compression: The Rate-distortion-perception Tradeoff[C]//International Conference on Machine Learning. Long Beach: PMLR, 2019: 675-685.

[20] LEDIG C, THEIS L, HUSZÁR F, et al. Photo-realistic Single Image Super-resolution Using a Generative Adversarial Network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4681-4690.

[21] ZHANG R, ISOLA P, EFROS A A, et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 586-595.

[22] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows[C]//2021 IEEE/CVF International Conference on Computer Vision(ICCV). Montreal: IEEE, 2021: 9992-10002.

[23] DING K Y, MA K D, WANG S Q, et al. Comparison of Full-reference Image Quality Models for Optimization of Image Processing Systems[J]. International Journal of Computer Vision, 2021, 129(4): 1258-1281.

[24] ZHOU W, SIMONCELLI E P, BOVIK A C. Multiscale Structural Similarity for Image Quality Assessment[C]//The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers. Pacific Grove: CIEEE, 2003: 1398-1402.

[25] ZHANG Z J. Improved Adam Optimizer for Deep Neural Networks[C]//Proceedings of 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Banff: IEEE, 2018: 1-2.

[26] THECKEDATH D, SEDAMKAR R R. Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks[J]. SN Computer Science, 2020, 1(2): 79.

作者简介:

伍忠东 男,(1968—),硕士,教授,硕士生导师。主要研究方向:深度学习、智能无线通信等。

甘炳坤 男,(2000—),硕士研究生。主要研究方向:语义通信等。

王鹏波 男,(2000—),硕士研究生。主要研究方向:深度学习、信号去噪及识别。

苟敬聪 男,(1998—),硕士研究生。主要研究方向:深度学习、信号去噪及分类。

丁尚思 男,(1999—),硕士研究生。主要研究方向:目标检测等。