面向机器视觉的文本提示引导的图像编码

黄志勐[1*],高 峰[2],杨 帆[2],马思伟[1]
(1. 北京大学计算机学院,北京 100871;2. 北京大学艺术学院,北京 100871)

摘 要: 近年来,随着物联网(Internet of Things,IoT)、语义通信以及智慧城市等经典机器间通信(Machine to Machine,M2M)场景的快速发展,海量视觉数据在设备间的实时传输与高效处理成为了一项关键挑战。在此背景下,传统以人眼感知质量为核心的图像编码方法,因其优化目标与机器视觉任务需求存在本质差异,往往在面向机器视觉分析时出现分析精度不足的问题。为此,面向机器视觉的图像编码(Image Coding for Machine,ICM)应运而生,其核心目标是在保证下游机器视觉任务(如分类、检测、分割等)分析精度的同时,实现尽可能低的编码码率,从而更好地适配M2M场景中的带宽与存储约束。然而,现有ICM方法仍面临两大瓶颈:其一,在极低码率条件下性能急剧下降。这是由于现有方法多依赖于端到端的非线性变换提取视觉特征,未能充分挖掘和利用图像中高层语义信息的紧凑表示,导致特征编码效率不足;其二,在开放场景下的泛化能力弱。多数方法针对单一任务、单一数据集进行优化,缺乏对未知类别、跨域数据的适应能力,难以在实际动态环境中保持稳定的分析性能。为突破上述限制,本文提出一种文本提示引导的面向机器视觉图像编码框架(Text-prompted Image Coding for Machine,T-ICM)。该框架的核心思想是将图像信息解耦为语义信息与纹理信息两个互补的组成部分,其中,语义信息以结构化文本提示(如对象类别、位置描述)的形式进行表示与编码,纹理信息则通过一种任务无关的通用视觉特征进行提取与压缩。在编码端,文本提示因其高度抽象和语义紧凑的特性,可以显著降低整体码率;通用特征则通过我们提出的分组特征编码模块进行高效压缩。在解码端,文本提示不仅用于直接解析完成分类、检测等任务,更重要的是作为引导信号,通过提示编码器与掩膜解码器,动态调整重建通用特征的语义感知区域,实现特征层面的域自适应与任务适配,从而显著提升模型在开放场景下的鲁棒性。本文在多个标准数据集与任务上对T-ICM进行了全面评估。实验表明,在语义分割和实例分割等密集预测任务上,T-ICM在极低码率下仍能保持接近原始图像输入的分析精度,其性能显著优于H.266/VVC、基于深度学习的图像编码器以及现有的其他ICM方法。本研究通过将语义信息迁移至高度压缩的文本模态进行传输,并利用其引导特征重建,T-ICM在编码效率与任务性能之间实现了更优的权衡,为未来语义通信、边缘智能协同,以及自适应机器视觉系统的发展提供了新的思路与技术支撑。

关键词: 视频编码;智能编码;特征编码;面向机器视觉的特征编码;深度学习;信号处理
基金项目: 国家自然科学基金(No.62025101,No.62176006);中国博士后科学基金(No.2025M771511)
中图分类号: TP37;TP39 **文献标识码:** A **文章编号:** 0372-2112(2026)01-0019-13
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250778

Text Prompted Image Coding for Machine

HUANG Zhimeng[1*], GAO Feng[2], YANG Fan[2], MA Siwei[1]
(1. *School of Computer Science*, *Peking University*, *Beijing* 100871, *China*;
2. *School of Art*, *Peking University*, *Beijing* 100871, *China*)

Abstract: In recent years, with the rapid development of classic machine-to-machine (M2M) communication scenarios such as the internet of things (IoT), semantic communication, and smart cities, the real-time transmission and efficient processing of massive visual data between devices have become a critical challenge. In this context, traditional image coding methods, which are primarily optimized for human perceptual quality, often suffer from insufficient analysis accuracy when applied to machine vision tasks due to a fundamental mismatch between their optimization objectives and the requirements of machine analysis. Consequently, image coding for machine (ICM) has emerged, aiming to maintain high analysis accuracy for downstream machine vision tasks (e.g., classification, detection, segmentation) while achieving the lowest possible bitrate, thereby better adapting to the bandwidth and storage constraints in M2M scenarios. However, existing ICM methods still face two major bottlenecks. First, their performance degrades sharply under extremely low bitrates. This is because most current approaches rely on end-to-end nonlinear transformations to extract visual features, failing to fully exploit the compact representation of high-level semantic information within images, which leads to inefficient feature coding. Sec-

收稿日期:2025-09-07;录用日期:2026-01-19;责任编辑:覃怀银
*通信作者:黄志勐

ond, they exhibit weak generalization in open-set scenarios. Most methods are optimized for single tasks or single datasets, lacking the adaptability to unseen categories or cross-domain data, and thus struggle to maintain stable analytical performance in practical, dynamic environments. To overcome these limitations, this paper proposes a novel text-prompted image coding for machine (T-ICM) framework. The core idea is to decouple image information into two complementary components: semantic information and texture information. The semantic information is represented and encoded in the form of structured text prompts (e.g., object categories, location descriptions), while the texture information is extracted and compressed as task-agnostic general visual features. At the encoder side, the text prompts, owing to their highly abstract and semantically compact nature, can significantly reduce the overall bitrate. The general features are efficiently compressed via our proposed grouped feature coding module. At the decoder side, the text prompts serve not only for direct parsing to accomplish tasks like classification and detection but, more importantly, act as guidance signals. Through a prompt encoder and a mask decoder, they dynamically adjust the semantically relevant regions of the reconstructed general features, enabling feature-level domain adaptation and task-specific adaptation, thereby significantly enhancing the model's robustness in open-set scenarios. The proposed T-ICM is comprehensively evaluated on multiple standard datasets and tasks. Experiments demonstrate that on dense prediction tasks such as semantic segmentation and instance segmentation, T-ICM can maintain analysis accuracy close to that of using the original uncompressed images even at very low bitrates, significantly outperforming H.266/VVC, learned image codecs, and other existing ICM methods. By migrating semantic information to the highly compressed text modality for transmission and utilizing it to guide feature reconstruction, T-ICM achieves a superior trade-off between coding efficiency and task performance. This work provides a novel perspective and technical foundation for the future development of semantic communication, collaborative edge intelligence, and adaptive machine vision systems.

Keywords: video coding; intelligent compression; feature coding; feature coding for machine; deep learning; signal processing

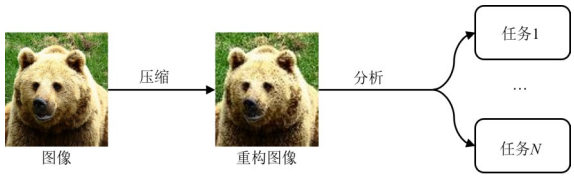
Foundation Item(s): National Natural Science Foundation of China (No.62025101,No.62176006); China Postdoctoral Science Foundation (No.2025M771511)

0 引言

近年来,互联网上图像内容呈现出指数级增长的趋势,推动着图像数据的传输与存储需求快速攀升,这一现实挑战催生了众多图像压缩技术的发展。图像编码的核心任务是将原始图像通过变换、量化、熵编码等模块压缩为紧凑的比特流,以提升存储效率并降低传输负担。长期以来,主流图像编码方法(如H.26x系列[1]、AVS系列[2]以及基于深度学习的端到端压缩模型[3-4])主要面向人眼观看,其优化目标侧重于最大程度保持信号保真度或感知质量。与此同时,深度学习技术的飞速发展也显著加速了各类机器视觉系统的落地,如物联网(Internet of Things,IoT)、语义通信与智慧城市等。这些系统所需的图像已从服务于人眼视觉逐渐转向为机器视觉分析。然而,受限于应用场景的不同,为人眼观看设计的图像编码方法在这些场景中往往会性能表现不佳。而在此背景下,面向机器视觉的图像编码(Image Coding for Machine,ICM)成为图像编码与计算机视觉交叉研究中的关键方向。

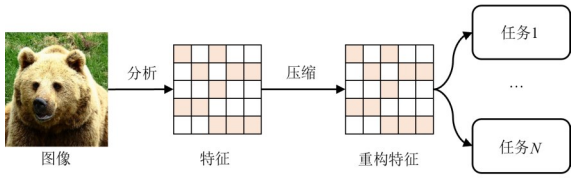
现有的ICM方法可以大致分为先压缩后分析(Compress-Then-Analysis,CTA)和先分析后压缩(Analysis-Then-Compress,ATC)两类范式[5]。遵循CTA范式(如图1(a)所示)的算法[6],首先对输入图像进行压缩,然后将重建图像输入至机器视觉算法中进行分析。但是同一张重建图像往往很难应对复杂的机器视觉任务场景。相对地,遵循ATC范式(如图1(b)所示)的方法[7],先在原始图像使用机器视觉算法进行分析,然后对分析得到的任务相关特征进行压缩。此类方法通常需要为每个特定任务生成独立的比特流,导致传输带宽需求显著增加。为降低传输开销,一些研究工作[8]尝试采用通用特征替代任务特异的特征进行编码,这在一定程度上缓解了多任务场景下的码率压力,但其码率效率仍然可以进一步提升。除了上述范式固有的问题以外,现有ICM方法仍普遍存在两大挑战:其一,编码效率低,难以实现极低码率压缩。这主要源于其编码过程(无论是对图像本身还是提取的特征)高度依赖于非线性变换模块。此类变换通常侧重于保持空间或通道维度的局部一致性,却未能充分挖掘和利用输入图像中蕴含的、对机器视觉任务更具本质性的高层语义信息的内在紧凑性,导致信息表示效率不高;其二,现有方法通常严重依赖于特定训练数据分布和预定义的标签空间,导致其在训练阶段未出现的目标类别或语义描述条件下难以适应,进而影响下游机器视觉任务性能的稳定性。为解决上述问题,本文提出了一种面向机器视觉的文本提示引导的图像编码方法

(Text prompted Image Coding for Machine, T-ICM)。该方法的核心思想是将图像信息解耦为语义信息和纹理信息,并分别表示为文本提示和通用特征。一方面,这种设计通过文本提示的紧凑编码有效解决了现有方法在超低码率下性能不佳的问题;另一方面,通过文本引导的特征自适应机制,有助于提升方法在分布外语义条件下的任务稳定性与适应能力。文本提示凭借其固有的语义紧凑性,能以极低的码率进行高效编码和传输。同时,在解码端,T-ICM利用接收到的文本提示来引导通用特征的重构与自适应调整,使其能够在不同下游任务或语义需求下进行自适应调整,从而提升方法在实际应用中的灵活性。通过引入跨模态建模的设计思想,T-ICM探索了ICM问题的一个新的研究思路,在兼顾高压缩率与强分析性能的同时,增强了在复杂语义条件下对机器视觉任务的适应能力。



(a) 先压缩后分析范式

(a) Compress-Then-Analysis



(b) 先分析后压缩范式

(b) Analysis-Then-Compress

图1　两种ICM算法的经典范式

Figure 1　Two typical paradigms of ICM algorithms

1　相关工作

1.1　图像编码算法

在传统研究中,图像编码的核心目标始终是优化人类视觉感知质量。大量方法通过信号保真度指标如峰值信噪比(Peak-to-Sinal-Noise-Rate, PSNR)、结构相似性(Structural SIMilarity, SSIM[9])或感知质量指标,如基于学习的感知图像块相似度指标(Learned Perceptual Image Patch Similarity, LPIPS[10])、恰可察失真(Just Noticeable Difference, JND[11])引导高质量重建。传统图像编解码器(如JPEG、H.26x、AVS)通常采用基于块的混合编码框架:首先,将图像分割为非重叠块,利用已重建像素对当前块进行预测(帧内/帧间预测);随后,通过离散余弦变换(Discrete Cosine Transform, DCT)、离散正弦变换(Discrete Sine Transform, DST)等算法将残差块转换至频域;变换系数经量化后,通过熵编码生成最终的比特流。近年来标准如AVS3[2]和H.266/VVC[1]引入样点自适应补偿(Sample Adaptive Offset, SAO[12])、自适应环路滤波(Adaptive Loop Filter, ALF[13])等多种滤波器以减弱块效应和振铃效应,进一步提升重建质量。

随着深度学习的发展,基于神经网络的图像压缩(Neural Image Compression, NIC)已成为图像编码领域的重要分支。主流方法以自编码器为基础架构[3],包括非线性变换网络、归一化模块与量化模块,并通过率失真联合优化方式进行训练。Ballé等人[14]引入了超先验模型,将隐变量建模为条件高斯分布,以提高熵估计精度;Cheng等人[15]进一步结合上下文建模与注意力机制,提出了基于混合高斯模型的压缩方法。近年来,研究者还探索了多种创新方向,以进一步提升压缩性能和语义保真能力。Lu等人[16]提出将Transformer结构引入图像压缩中,以建模长距离依赖并改善图像整体结构重建;Li等人[17]在ICLR2024提出Frequency-Aware Transformer(FAT),在频域上对不同频段进行建模以适配多种图像结构。此外,Diffusion-based压缩模型也逐步成为研究热点,Theis等人[18]提出的DiffC方法与Xia等人[19]的DiffPC模型,展示了扩散模型在极低码率下保持良好感知质量的潜力。同时,分层可伸缩压缩架构(如Ballé的Scale Hyperprior模型)实现了码率灵活调控,为多终端场景提供支持;M2T[20]等工作则通过二次掩码提升解码效率。上述工作从架构、建模和多目标优化角度不断丰富图像压缩技术体系,为本文提出的特征压缩模块提供了有力的技术启发。然而,上述方法主要围绕人眼视觉质量进行优化,其设计目标与面向机器视觉任务的编码需求仍存在本质差异。

1.2　面向机器视觉的图像编码算法

随着机器视觉在物联网、安防、智慧城市等领域中的广泛应用,面向机器视觉的图像编码研究成为近年来的前沿方向。早期部分研究借鉴传统面向人眼的感兴趣区域(Region Of Interest, ROI)编码理念,在图像中提取人眼关注区域并给予更高码率以提升分析性能[21]。但机器视觉的关注区域与人眼存在本质差异,因此后续研究提出了面向机器的感兴趣区域(Region Of Interest for Machine, ROIM)概念[6],以机器感兴趣区域为目标分配码率,能够提升编码图像在机器视觉任务上的分析性能。

在基于学习的图像编码方法的基础上,引入机器视觉任务特定的损失函数可以有效提升下游机器视觉分析性能[22-23]。Patwa等人[24]在端到端压缩结构

中引入额外的卷积层和 softmax 层来计算分类损失，与压缩损失联合优化，有效提升了其在分类任务上的分析准确率；Chamain 等人[25]使用了一个针对目标检测任务的损失函数来直接优化端到端图像编码器，有效提升了在目标检测任务上的性能；Le 等人[26]提出了一种内容自适应的端到端图像编码方法，来支持实例分割任务；SSSIC 方法[27]实现了从语义到信号的多级可伸缩压缩，以适配不同类型的下游任务。然而，此类方法大多依赖特定任务与数据集进行联合优化，导致其在训练阶段未出现的目标类别或语义条件下，难以稳定支撑多任务分析性能。为此，一些研究引入辅助特征码流，以提升分析鲁棒性与通用性。Feng 等人[28]提出的 Omni-ICM 框架提取通用特征，用于多任务分析；P-ICM 方法[8]在此基础上设计了任务驱动的 prompt 机制，用于特征调控；TransTIC 方法[29]则进一步引入 Transformer 结构，以从人眼感知迁移到机器感知领域。上述方法均致力于缓解 ICM 在任务切换与跨域场景下的性能退化问题。需要指出的是，上述方法中引入的 prompt 或调控机制，主要用于调节编码器或特征提取过程，其语义信息通常以隐式向量形式存在，并未作为独立的信息表示参与编码与传输过程。相应地，语义信息对码率分配和解码重建的影响更多体现在网络参数层面，而缺乏显式可控的表达形式。

此外，近期部分工作开始探索更高层的语义模态作为压缩信息传输形式。Xia 等人[30]和 Chen 等人[29]分别在图像生成和机器感知中引入文本或语言模态以表达语义信息。相比传统图像/特征表示，文本模态具备更强的语义压缩能力与引导能力，为解决 ICM 的低码率与跨任务挑战提供了全新路径。本文提出的 T-ICM 方法正是基于上述思考，通过将图像语义信息抽象为文本提示参与编码与引导，并辅以通用特征流进行自适应解码，以探索在 ICM 场景下兼顾编码效率与任务适应性的一种可行路径。同时，与 P-ICM 和 TransTIC 等方法不同，本文提示机制并不干预编码器参数，而是直接用于重建特征的语义引导，在机制层面上实现了语义信息与纹理特征的显式解耦，为后续编码与解码过程中的灵活调控提供了更大的设计空间。

2 T-ICM 算法框架与实现

在 2.1 节介绍整体框架设计思路与核心优化目标，重点解释语义-纹理解耦与文本引导机制。在第 2.2~2.6 节中，分别对各主要模块的实现细节进行说明，包括通用特征提取器、提示词生成器、图像编码器、提示词编码器以及任务解码器。通过这一整体到局部的介绍，系统展示 T-ICM 的方法设计与实现路径。

2.1 整体框架与信息流

T-ICM 的编解码流程如图 2 所示，其核心是通过语义-纹理解耦与文本动态引导，实现面向机器视觉的高效编码。

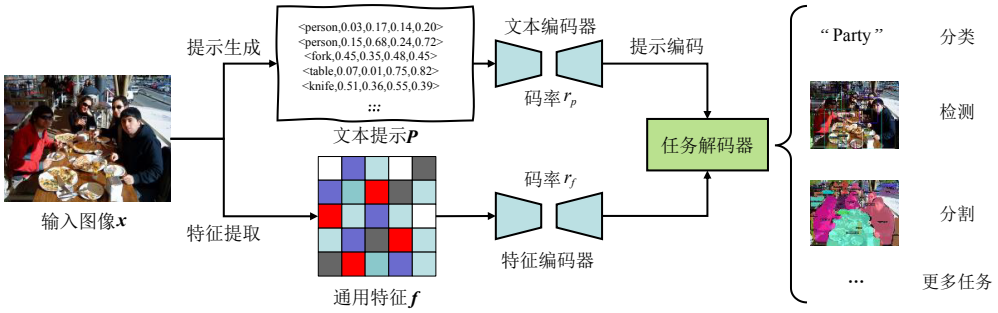


图 2 算法框架图

Figure 2 Framework of the proposed algorithm

在该框架中，文本提示并非仅作为推理阶段的辅助引导信息，而是作为一种显式的语义表示，与通用特征一同参与编码、传输与解码过程。对于输入图像 $\boldsymbol{x}$ 采用双路处理：一方面，经由特征提取器（Feature Extractor，FE）得到通用特征 $\boldsymbol{f}$；另一方面，通过提示词生成器（Prompt Generator）生成文本提示 $\boldsymbol{p}$：

$$\boldsymbol{f}=\mathrm{FE}\left(\boldsymbol{x}\middle|\theta_{\mathrm{FE}}^{*}\right) \tag{1}$$

$$\boldsymbol{p}=\left\{(\boldsymbol{c}_i,\boldsymbol{b}_i)\right\}(i=1,2,\cdots,K)=\mathrm{PG}(\boldsymbol{x}) \tag{2}$$

其中，θ_{FE}^{*} 为特征提取器中的可学习参数，* 符号表示该参数在训练过程中是固定的。文本提示 $\boldsymbol{p}$ 作为压缩过程的另一关键组件，由文本对 $(\boldsymbol{c}_i,\boldsymbol{b}_i)$ 构成。此处 $(\boldsymbol{c}_i,\boldsymbol{b}_i)$ 分别表示第 i 个对象的类别与位置。文本提示的获取方式具有灵活性，既可以在解码端由用户自定义输入，也可以由现有目标检测算法自动提取。需要注意的是，$\boldsymbol{c}_i$ 和 $\boldsymbol{b}_i$ 两者只存在其中之一就可以进行推理，不需要两部分信息都存在。下一步是将通用特征 $\boldsymbol{f}$ 和文本提示 $\boldsymbol{p}$ 编码成比特流，编码过程可以形式化为

$$r_p, \boldsymbol{b}_p = E_p(\boldsymbol{p}) \tag{3}$$

$$r_f, \boldsymbol{b}_f = E_f\left(\boldsymbol{f}|\theta_{E_f}\right) \tag{4}$$

其中,r_p和r_f代表$\boldsymbol{p}$和$\boldsymbol{f}$对应的编码器估计出的编码所需比特数;$\boldsymbol{b}_p$和$\boldsymbol{b}_f$表示为存储和传输生成的比特流;θ_{E_f}表示E_f模块中训练时可学习的参数。然后将编码后的比特流输入到相应的解码器中,以重构文本提示词和特征。这一步可以表述为

$$\hat{\boldsymbol{p}} = D_p(\boldsymbol{b}_p) \tag{5}$$

$$\hat{\boldsymbol{f}} = D_f\left(\boldsymbol{b}_f|\theta_{D_f}\right) \tag{6}$$

其中,$\hat{\boldsymbol{f}}$表示重建特征;θ_{D_f}代表解码器D_f的学习参数,为重建文本提示。值得注意的是,E_p和D_p构成的文本编解码器为无损编码器,这样可以确保重构文本$\hat{\boldsymbol{p}}$与原始文本$\boldsymbol{p}$完全一致。

随后,解码特征$\hat{\boldsymbol{f}}$与解码文本$\hat{\boldsymbol{p}}$被送入任务解码器(Task Decoder,TD),用于各类机器视觉任务的分析。根据不同机器视觉任务的特点,本文将常见的机器视觉任务划分为密集预测任务和对象级任务两大类别。针对语义分割与实例分割等密集预测任务,首先通过提示编码器(Prompt Encoder,PE)将文本转化为更易理解的特征表示$\tilde{\boldsymbol{p}}$,该转换过程可表述为

$$\tilde{\boldsymbol{p}} = \mathrm{PE}\left(\hat{\boldsymbol{p}}|\theta_{\mathrm{PE}}\right) \tag{7}$$

其中,θ_{PE}表示提示编码器中的可学习参数。通过将变换后的文本提示$\tilde{\boldsymbol{p}}$与重建特征$\hat{\boldsymbol{f}}$融合送入任务解码器中生成具有语义信息的掩膜$\boldsymbol{m}$,该操作可表述为

$$\boldsymbol{m} = \mathrm{TD}\left(\tilde{\boldsymbol{p}}, \hat{\boldsymbol{f}}|\theta_{\mathrm{TD}}\right) \tag{8}$$

其中,θ_{TD}表示任务解码器的可学习参数。由于输入图像的重建特征在固定码率下保持不变,系统可通过文本提示动态生成注意力掩膜,从而聚焦于不同目标区域。考虑到现有ICM方法普遍支持对象级任务(如分类与检测),T-ICM框架同样保留了这一能力,通过直接解析解码后的文本提示$\left\{(\boldsymbol{c}_i, \boldsymbol{b}_i)\right\}_{i=1,2,\cdots,K}$即可完成对象级任务分析。

2.2　特征提取器

为提升算法的泛化能力,本文提出了一种任务无关的通用特征表示,该表示需要同时满足两个要求:一是具备在任意位置生成语义掩膜的空间灵活性;二是能够在不同数据域间保持高层语义的一致性。基于上述需求,本文采用Mask AutoEncoder(MAE)[31]预训练的ViT-H/16[32]作为基础架构。其自监督预训练机制为特征学习提供了良好的泛化能力,而分层注意力结构则有助于有效融合局部与全局的上下文信息。为了进一步适配高分辨率输入,对原始架构进行了改进:一是引入14 × 14窗口注意力机制以降低计算复杂度;二是在每隔四层插入全局注意力模块,以保持长程依赖的建模能力。最终,特征提取器能够将输入图像$\boldsymbol{x}$映射为通用特征$\boldsymbol{f}$,为后续的编码传输、提示引导与任务解码提供稳定的表示基础。需要指出的是,该特征提取器的选择主要基于其稳定的语义表示能力,T-ICM框架本身并不依赖于特定模型规模,其他不同容量的视觉编码器亦可在保持整体结构不变的情况下进行替换。

2.3　提示词生成器

在T-ICM中,提示词生成器采用形如〈类别,边界框〉的文本对作为机器视觉任务的语义引导信号,其生成方式既可以依托现有目标检测模型(如YOLO[33]、DINO[34]等)自动获得目标类别与位置,也可以根据具体应用场景由用户手动指定目标描述。考虑到无损压缩的码率约束,所有提示信息均需转换为纯文本格式,例如边界框编码为归一化坐标文本"0.25,0.33,0.75,0.82"。此外,大语言模型(如Woodpecker[35])可作为增强方案生成细粒度语义描述,如"十字路口左侧的出租车",该方案通过视觉语义对齐提升提示质量(具体案例参见3.5节),但同时也会带来额外的计算资源开销。

2.4　编码器

文本编码器。文本与特征在T-ICM中采用独立的编码策略。由于文本提示的语义高度敏感,即使极小的偏差也可能导致下游机器视觉任务失败,因此本文对文本提示$\boldsymbol{p}$采用无损压缩。具体实现上使用哈夫曼编码(Huffman Coding),其码率可形式化为

$$r_p = -\sum_{i=1}^{N} \log_2\left(P\left(v_p(i)\right)\right) \tag{9}$$

其中,P表示符号$v_p(i)$在词表中的出现频率;r_p为编码文本$\boldsymbol{p}$所需的总比特数。该方法能够确保文本提示在压缩与解码后保持严格一致,从而保证语义信息的可靠传输。

特征编码器。在特征编码方面,本文提出一种基于深度学习分组特征编码模块(Grouped Feature Coding,GFC),其网络架构如图3所示。整体流程分为四步:首先将特征$\boldsymbol{f}$经填充层扩展通道维度为258,随后均匀为k个子组以提升压缩效率,各组经共享权重的编码网络[15]进行编码。编码完成后,通过聚合层融合所有子特征组以形成完整的比特流。其中,原始通用特征的通道维度为256。为适配后续分组编码结构,本文对特征维度进行最小化补全,使其可被子编码器的输入通道数整除。所采用的子编码器结构继承自成熟的图像压缩网络,其基本处理单元为3通道输入,该设计在参数共享、结构稳定性与工程迁移性方面具有良好实践基础。在此设置下,通道维度补全至

258后可自然划分为86个子组，子组数量由结构约束确定，而非通过性能调参获得。

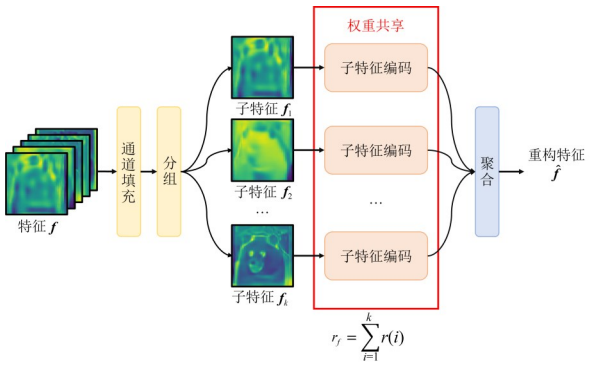


图3 特征编码器框架图

Figure 3 Architecture of the feature codec

在上述分组编码框架中，每个子组所采用的共享权重编码网络继承自Cheng等人[15]提出的基于超先验的深度学习图像压缩模型。该子编码器由分析变换网络、合成变换网络以及超先验分支组成，其结构在各子组之间完全共享。具体而言，分析变换网络由四个卷积层级联构成，每一层采用5×5卷积核、步长为2，并结合非线性激活函数，用于逐级降低空间分辨率，并将3通道子特征映射至潜在表示空间。在潜在表示空间中，引入超先验分支对其统计分布进行建模，通过额外的超分析与超合成变换预测潜在表示的条件概率分布，以提升熵编码的精度。合成变换网络在结构上与分析变换网络对称，同样由四个采用5×5卷积核、步长为2的反卷积层构成，用于逐级恢复子特征的空间分辨率。通过对所有子组编码结果进行聚合，最终形成完整的特征比特流。

训练过程中，针对不可导的量化过程，在潜在表示中引入加性均匀噪声以近似量化操作来实现码率估计，该过程可形式化表述为

$$\begin{aligned} L_{\mathrm{fc}} &= R+\lambda D \\ &= E\left(-\log_2\left(p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}})\right)\right)+E\left(-\log_2\left(p_{\hat{\boldsymbol{z}}|\boldsymbol{\varphi}}(\hat{\boldsymbol{z}}|\boldsymbol{\varphi})\right)\right)+\lambda d(\boldsymbol{f},\hat{\boldsymbol{f}}) \end{aligned} \tag{10}$$

其中，$\hat{\boldsymbol{y}}$表示由E_f模块编码的隐表示；$\hat{\boldsymbol{z}}$代表$\hat{\boldsymbol{y}}$的超先验信息；概率分布$p_{\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}}}(\hat{\boldsymbol{y}}|\hat{\boldsymbol{z}})$与$p_{\hat{\boldsymbol{z}}|\boldsymbol{\varphi}}(\hat{\boldsymbol{z}}|\boldsymbol{\varphi})$分别描述$\hat{\boldsymbol{y}}$和$\hat{\boldsymbol{z}}$的统计特性；$\boldsymbol{\varphi}$则是编码$\hat{\boldsymbol{z}}$时候的可因式分解模型。函数$d$的定义为原始特征与重建特征间的均方误差（Mean Square Error，MSE），其计算公式为

$$d(\boldsymbol{f},\hat{\boldsymbol{f}})=\sum_{h=1}^{H_f}\sum_{w=1}^{W_f}\frac{(\boldsymbol{f}_{h,w}-\hat{\boldsymbol{f}}_{h,w})^2}{H_fW_f} \tag{11}$$

2.5 提示词编码器

提示词编码器的输入为形式化后的边界框位置信息，其格式遵循COCO数据集的标注规范。每个边界框可表示为一对嵌入：第一分量由边界框左上角点的位置编码与一个可学习的嵌入向量相加；第二分量则由右下角点的位置编码与另一可学习嵌入向量相加，从而获得兼具位置约束与可学习性的表示。在完成上述形式化后，系统采用CLIP模型[36]对提示词进行高维语义特征提取。由于CLIP经过大规模视觉-语言对齐预训练，该设计能够显著增强边界框位置信息的语义表达能力，从而在下游任务中更有效地引导特征解码。

2.6 掩膜解码器

本文采用SAM模型[37]的轻量化掩码解码器作为任务解码器的基础架构，该解码器由提示嵌入模块、特征融合模块以及多层感知器掩膜预测头构成。然而，由于特征编解码过程中不可避免地失真，直接迁移预训练模型往往导致性能显著下降。为缓解这一问题，本文在解码器上进一步引入多目标损失函数进行微调，其优化目标可表示为

$$L_2=\alpha L_{\mathrm{focal}}+L_{\mathrm{dice}}+L_{\mathrm{iou}} \tag{12}$$

其中，焦点损失L_{focal}定义为

$$L_{\mathrm{focal}}=-\beta(1-p_t)^{\gamma}\log p_t \tag{13}$$

其中，p_t表示正确类别的预测概率。参照文献[38]，超参数α、β和γ分别设为20、0.8和2。Dice损失L_{dice}的计算公式为

$$L_{\mathrm{dice}}=1-\frac{2|\boldsymbol{M}_p\cap\boldsymbol{M}_g|+s}{|\boldsymbol{M}_p|+|\boldsymbol{M}_g|+s} \tag{14}$$

其中，$\boldsymbol{M}_p$和$\boldsymbol{M}_g$分别表示预测掩码与真值掩码，平滑项$s=1$用于防止除零错误。交并比损失L_{iou}定义为交并比预测值与预测掩码-真值掩码实际交并比间的均方误差。通过上述多目标联合优化，掩码解码器能够在保持语义一致性的同时，有效缓解特征失真带来的性能退化。

3 实验结果

3.1 对比方法与实验配置

3.1.1 对比方法

为了全面评估T-ICM的性能，本文选取了四类方法作为对比对象：传统图像编码方法（HEVC、VVC）、基于深度学习的图像编码方法（cheng2020[15]）、面向机器视觉的图像编码方法（Prompt-ICM[8]）和分析模型基线。对于传统编码方法和基于深度学习的编码方法，实验过程中首先对输入图像进行编码得到码率和重构图像，然后将这些重构图像输入到机器视觉算法中绘制码率-准确率性能曲线。Prompt-ICM的源代码尚未开源，本文直接采用其论文中报告的实验结果。分析模型基线是为了计算未压缩场景下机器视觉算法的理论上限，将原始图像直接输入机器视觉算

法得到。

此外,近年来亦有研究尝试在图像编码或压缩框架中引入文本或语义标签作为提示信息,以引导生成或重建过程,从而提升语义一致性或感知质量。代表性方法如Diff-ICMH[39],通过将图像级语义标签作为条件信息,引导生成式压缩模型完成图像重建。需要指出的是,该类方法中的文本提示主要作为生成阶段的条件约束,并未作为需要显式编码与传输的语义信息载体参与比特分配,其研究目标亦主要聚焦于视觉重建质量。鉴于其问题设定与本文面向机器视觉任务性能优化的ICM框架存在差异,本文未将该类方法纳入主对比实验,而是在后续实验中对其进行补充分析,用以比较不同文本介入方式在低码率条件下对机器视觉任务性能的影响。

3.1.2　数据集

在训练阶段,T-ICM根据不同模块的需求使用了多种数据集。首先,使用SA-1B数据集[37]对特征提取器与任务解码器进行训练。该数据集包含超过1 100万张高分辨率图像及约11亿个紧凑掩码标注,能够有效支持模型的通用特征学习。其次,采用MS COCO 2017数据集[40]对特征编码器进行训练并对任务解码器进行微调。该数据集包含118 000张训练图像与5 000张验证图像,是目标检测与语义分割的经典基准。

在测试阶段,T-ICM在图像分类、目标检测、语义分割和实例分割四个不同的机器视觉任务上开展实验。其中,MSCOCO 2017用于评估目标检测与语义分割性能;TrashCan-1.0[41]用于验证实例分割任务,检验模型在特定领域的适应性;CUB-200-2011[42]、Stanford-Dog[43]与Stanford-Car数据集则用于图像分类任务,进一步说明所提方法在简单视觉任务中的兼容性。

3.1.3　评价指标

为了测试T-ICM的编码分析效率,本文绘制了所提方法在各种机器视觉任务上的码率-分析效率曲线。码率和传统编码的计算方法一致,计算公式为

$$\mathrm{bpp}=\frac{r_f+r_p}{hw} \tag{15}$$

其中,w和h表示输入图像的宽度和高度。在性能度量方面,对于需要同时使用文本提示与通用特征的任务(语义分割与实例分割),本文采用平均精度均值(mean Average Precision,mAP)作为核心指标,并将mAP替代传统BD-Rate[44]算法中的PSNR,以计算所提方法节省的码率。对于图像分类与目标检测任务,由于可直接通过解析文本提示进行分析,分别采用Top-1 Accuracy与mAP进行评价。由于这两类任务的性能与码率无关,其码率-分析效率曲线为平行于横轴的直线。为简化展示,仅在性能图中绘制最低比特率对应的结果点。

3.2　训练策略

整个训练流程分为三个阶段,训练阶段一为预训练阶段,训练阶段二和训练阶段三中各个模块的状态如表1所示。训练流程中模块的状态共有三种:更新(Update)、冻结(Freeze)和跳过(Skip)。其中,更新表示模块的参数在反向传播中根据损失函数进行优化;冻结表示模块参与前向推理,但其参数在训练过程中保持不变;跳过则指模块在训练过程中不被执行,其输入直接传递至下一个模块。在上述三个阶段训练流程中,实际涉及参数更新的核心模块包括特征编码器和任务解码器。其中,特征编码器采用基于超先验的卷积编码结构;任务解码器基于SAM模型中的prompt-driven mask decoder架构,由提示嵌入、特征融合及掩膜预测模块组成。其余模块(如特征提取器与提示词编码器)在训练过程中作为固定特征变换或语义对齐组件参与前向推理,不涉及参数更新。在预训练阶段(阶段一),本文采用ViT-H/16、CLIP以及SAM掩码解码器的预训练模型作为初始化参数,并使用SA-1B数据集对除特征编码器外的模块进行预训练。在训练阶段二,T-ICM将特征提取模块的可学习参数保持冻结,仅以式(13)所定义的率失真损失函数优化特征编解码器。对应的λ的数值分别为[0.001 8, 0.003 5, 0.006 7, 0.013, 0.025, 0.048 3]。优化的学习率设为5×10^{-5},训练2个epoch,在此阶段,掩码解码器与提示编码器保持跳过状态,训练数据集为MSCOCO 2017与SA-1B。因此,在训练阶段二中,仅对特征编码器的变换、反变换及超先验分支参数进行更新,其余模块均保持冻结或跳过状态。在训练阶段三,以式(14)所定义的损失函数对掩码解码器进行优化,而图像编码器与提示编码器保持冻结,文本提示编码器继续保持跳过状态。针对阶段两种不同码率的模型,以学习率1×10^{-4}进行20个epoch的优化。在训练阶段三中,仅对任务解码器中的提示融合与掩膜预测相关参数进行优化,以增强其在压缩特征输入条件下的鲁棒性。在阶段三中,为提升训练效率,将训练集图像预处理为对应的特征数据库以供调用。所有实验在NVIDIA RTX 4090平台上完成,数据增强策略遵循文献[45]中提出的增强方法。

为进一步说明训练过程中超参数的设定依据,本文对所涉及参数类型进行如下区分说明:一类为与编码结构直接相关的结构参数,例如特征分组方式及子组数量等,其取值由编码网络结构及分组约束自然确定,而非通过性能调参获得,该类参数在模型结构确

表1 训练过程中每一个模块的参数更新状态

Table 1 Status of each module during training process

模块	训练阶段二	训练阶段三
文本编码器&提示词生成器	跳过	跳过
特征提取器&提示词编码器	冻结	冻结
特征编码器	更新	冻结
掩膜解码器	冻结	更新

定后保持固定;另一类为训练配置参数,包括率失真权重λ、学习率及训练轮数等,其取值主要参考相关图像压缩与机器视觉任务中的常见设置,并结合模型在训练过程中的收敛性与稳定性进行确定。实验中采用不同λ值以覆盖多个码率点,其余训练参数在保证稳定收敛的前提下保持一致。

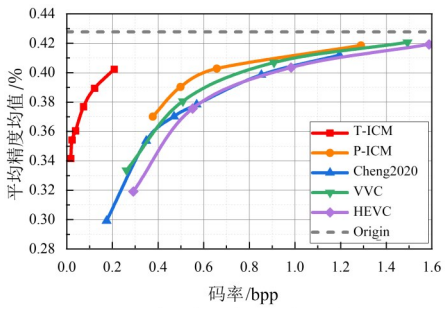
3.3 实验结果

在语义分割任务上,本文在MSCOCO 2017数据集上使用mmdetection框架实现的Mask R-CNN模型进行验证,模型采用官方预训练参数。实验结果如图4(a)所示,T-ICM在极低码率(约0.017 bpp)下仍能达到0.34的mAP,相较未压缩输入仅下降约0.08,显示出在高压缩比条件下仍具备可接受的精度表现。相比之下,传统方法与基于学习的压缩方法需在10~20倍更高码率下才能达到相近性能。由于P-ICM的论文未提供低码率结果,本文无法在最低码率点进行直接对比,但从第四个码率点结果可见,T-ICM在相似精度下码率显著低于P-ICM。

在实例分割任务上,本文在TrashCan-1.0数据集上使用mmdetection框架的Mask R-CNN算法进行验证,并按照官方流程在TrashCan-1.0训练集上微调模型。为了检验T-ICM的任务适配性,本文直接使用语义分割任务中训练好的模型,仅修改提示词生成器以对齐标签。实验结果如图4(b)所示,T-ICM在压缩倍率和同码率条件下均显著优于对比方法。

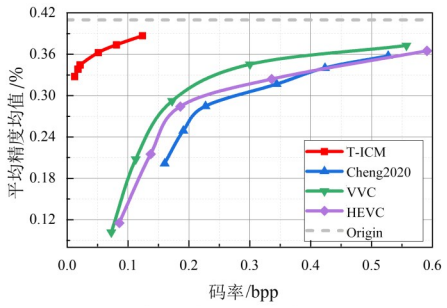
在图像分类任务上,本文在Stanford-Dog、Stanford-Car和CUB-200-2011数据集上进行评测,测试过程遵循文献[7]中描述的方法,输入图像缩放至224 × 224。由于图像分类任务可直接通过文本解析获得结果,不受特征码流变化影响,其码率-性能曲线理论上为平行于X轴的直线。为简化展示,本文仅在图中标注最低码率点(红点)。最终结果如图5(b)、图5(c)和图5(d)所示,T-ICM在性能-码率坐标系中始终位于左上区域,即在保证较低码率的同时取得了优于对比方法的性能。

在目标检测任务上,本文同样在MSCOCO 2017数据集上使用Mask R-CNN模型进行评测。由于该任务与分类类似,仅需检测框和类别而无需像素级掩码,因此可直接通过文本码流解析获得结果。如图5(a)所示,T-ICM在极低码率条件下即可超过对比方法,表现出较高的效率与适应性。



(a) 语义分割任务上的性能曲线

(a) Performance on semantic segmentation



(b) 实例分割任务上的性能曲线

(b) Performance on instance segmentation

图4 语义分割和实例分割任务上的性能对比曲线图

Figure 4 Rate-performance comparison on semantic segmentation and instance segmentation

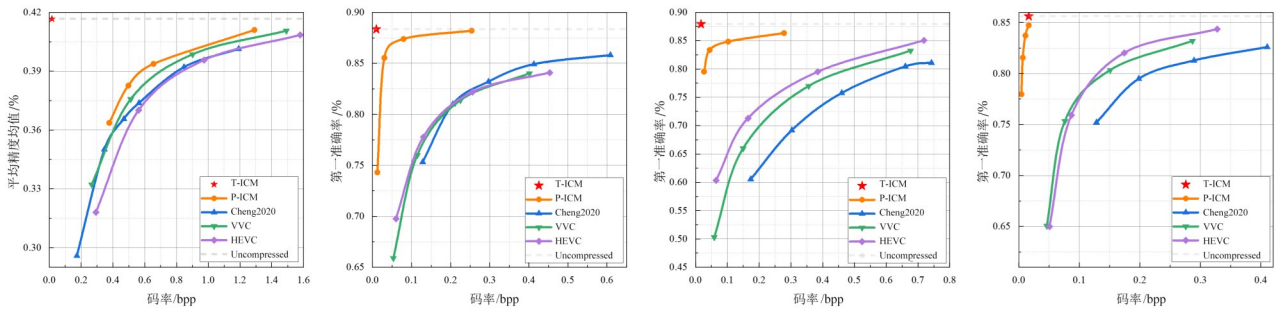
3.4 文本提示条件的生成式图像编码方法对比

本节对第3.1.1节中引入的文本提示条件图像编码方法Diff-ICMH[39]在MS COCO数据集上的性能进行对比。实验统一在相同的数据集划分、评价模型及码率计算方式下进行,同时评估目标检测与语义分割的分析性能,相关结果汇总于表2。

从表2可以观察到,在同时覆盖目标检测与分割任务的统一对比中,本文提出的T-ICM方法在更低的码率条件下取得了更高的任务分析性能。相比之下,文本提示条件的生成式图像编码方法在相同或更高码率下,其检测与分割性能均存在明显差距。该结果表明,在COCO数据集的复杂视觉任务场景中,T-ICM具备更优的整体表现。

3.5 特征编码模块选择

基于深度学习的图像编码框架通常是面向3通道彩色图像设计的,很难直接迁移到多通道的特征上,因此必须要设计特征维度改变策略才能实现高效编码。本文提出了一个基于分组压缩的特征编码模块,这个模块可以有效地降低模型参数,简化训练流程,提升编码效率。本文设计了一组消融实验来



(a) MSCOCO 2017/目标检测　(b) CUB/图像分类　(c) Standford-Dog/图像分类　(d) Standford-Car/图像分类

(a) MSCOCO 2017/object detection　(b) CUB/image classification　(c) Standford-Dog/image classification　(d) Standford-Car/image classification

图5　目标检测和图像分类任务上的性能对比曲线图

Figure 5　Rate-performance comparison on object detection and image classification

表2　Diff-ICMH与T-ICM性能对比

Table 2　Comparison between Diff-ICMH and T-ICM

测试方法	码率	目标检测性能	语义分割性能
Diff-ICMH	0.013	21.26	18.19
	0.032	27.94	24.91
	0.068	32.22	29.10
	0.104	34.50	31.80
T-ICM(Ours)	0.017	41.67	34.16
	0.038	41.67	36.04
	0.073	41.67	37.68
	0.122	41.67	38.94

对比T-ICM中分组特征编码模块在MSCOCO 2017数据集上的性能。对照算法包括使用卷积进行维度改变(Transform)和直接修改模型输入的通道层数(Modify)两种方法,实验结果如图6所示,评价指标为*F*-分数。从折线图中看到T-ICM中所设计的分组编码模块性能远好于另外两种编码策略:Transform方法可以有效降低特征码率,但是卷积改变维度过程中信息损失较大,分析性能下降较多;Modify方法一定程度上可以保持分析性能,但是由于没有降低特征数目,码率相对较高,不能够实现极低码率下的特征编码。

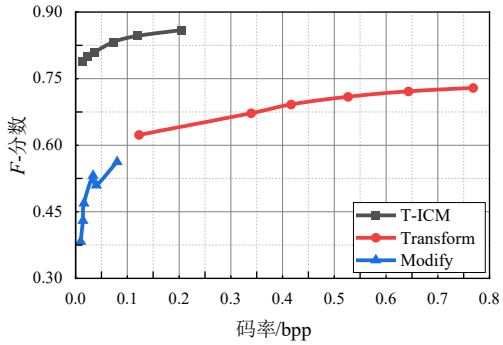


图6　不同特征编码模块的对比实验

Figure 6　Ablation study between different feature codec

3.6　特征提取模块配置

在基于预训练的通用特征模型中,模型规模是影响性能的关键因素之一。为此,T-ICM中选择了参数量最大的ViT-H模型作为特征提取器。为探究模型尺度对编码性能的影响,本节在MSCOCO 2017数据集上对比了不同规模的ViT模型(包括ViT-L和ViT-B)。为量化性能差异,实验以T-ICM(ViT-H)为基准,采用BDRate指标评估ViT-L和ViT-B版本在平均精度均值(mAP)上的性能损失,具体结果如表3所示。实验表明,模型规模对编码性能会产生影响,编码性能随着模型的增大而提升。特征提取器的模型容量会对编码性能产生影响,在当前设置下,ViT-H取得了更优的任务性能-码率折中。进一步提升特征提取模块性能仍可能带来收益,但其效果与具体模型结构和训练配置相关,有待进一步系统分析。

表3　不同特征提取器之间的性能对比

Table 3　Comparison between different feature extractors

模型	性能对比
ViT-H	0
ViT-L	5.68%
ViT-B	6.12%

3.7　提示词生成器讨论

提示词生成器是一个训练无关的可替换模块,可替换的内容包括视觉分析模型(Visual Analysis Model,VAM)、用户自定义(USR)和多模态大语言模型(Multimodel Large Language Models,MLLM)。本节将会给出一些实例来用于展开介绍使用不同提示词生成器的T-ICM方法,给出的所有例子的码率均为0.07 bpp左右,该码率下大部分面向机器视觉的图像编码方法的分析性能均比较低,语义理解能力较差,不具备实用性。除定性案例外,我们进一步补充了不同提示词生成器在未压缩/压缩两种条件下的定量对

比,以分析提示质量上限与低码率语义保留能力之间的关系。

T-ICM(VAM)。与视觉分析模型相结合的T-ICM方法具有最广泛的适用范围,能够应用于大多数具备明确分析任务的机器视觉应用场景。以使用Faster-rcnn模型作为视觉分析模型为例:首先将原始图像输入Faster-rcnn,获取一组类别坐标对作为文本提示词;随后将这些提示词送入提示词编码器进行特征提取;接着利用编码后的文本引导解码特征;最终得到语义分析结果。整个流程图如图7所示,整个过程中编码码率为0.07左右,实现了低码率下的面向机器视觉的编码。

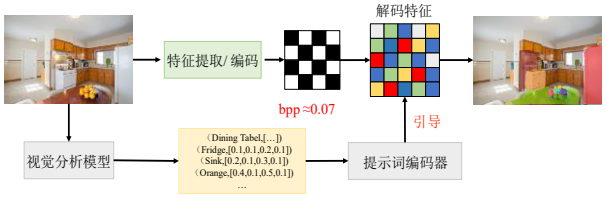


图7 不同特征编码模块的对比实验

Figure 7 Ablation study between different feature codec

T-ICM(USR)。基于用户自定义的提示词生成器主要分为自定义类别和自定义区域两类。自定义类别是指提示词生成器的输入为任意类别文本,然后通过Owl-ViT[46]或Grounding DINO[47]等目标检测标注框生成类别坐标对,送入提示词编码器进行编码、引导和解码;自定义区域的方法则更为简洁,直接在编码端根据需求选择需解析的区域,将选择区域作为提示词编码器的输入进行后续处理。这两类方法充分展现了T-ICM在开放集合标签上的鲁棒性,整个流程简图如图8所示。输入原图在MSCOCO 2017数据集中并未包含标签“门”的标注,T-ICM(USR)通过自行输入标签“门”或选择区域进行分析、编码与传输,有效地根据具体现实应用场景实现了面向机器视觉的编码。

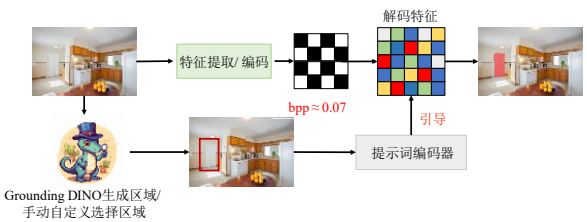


图8 不同特征编码模块的对比实验

Figure 8 Ablation study between different feature codec

T-ICM(MLLM)。多模态大语言模型具备高效对齐图像与语义的能力,这一特性使其十分适合应用于提示词生成器模块。本文以Woodpecker[34]为例,构建了基于多模态大语言模型的T-ICM方法。Woodpecker通过构建一系列提示,能够生成附带坐标位置的文本描述,该算法最初旨在解决多模态大语言模型中的幻觉问题。本文利用这一特性,将其作为T-ICM的提示词生成器,在解码端解析其中的物体坐标对,以用于场景的语义理解,可以在较低码率下实现高效的面向机器视觉的编码。具体框架流程如图9所示。

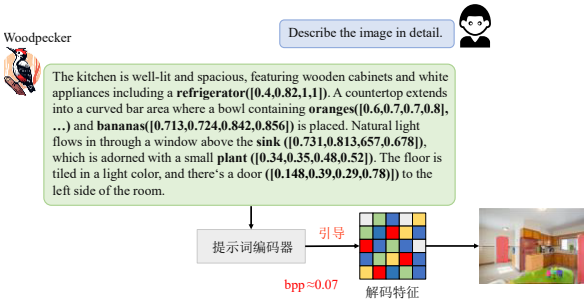


图9 不同特征编码模块的对比实验

Figure 9 Ablation study between different feature codec

为进一步分析不同提示词生成方式对下游任务性能的影响,本文在COCO数据集上对多种提示词生成策略进行了定量对比分析,结果如表4所示。

表4 不同提示词生成方式在COCO数据集上的定量性能比较

Table 4 Comparison between different feature extractors

测试方法	压缩状态	码率	目标检测性能	语义分割性能
VAM	压缩后	0.122	41.67	38.94
	未压缩	—	41.67	42.87
MLLM	压缩后	0.123	49.67	45.28
	未压缩	—	49.67	47.12

实验结果表明,不同提示词生成方式在未压缩条件下的任务性能存在一定差异,反映了语义提示质量对分析性能上限的影响。在低码率压缩条件下,虽然提示词质量仍对性能产生影响,但其差异并不与未压缩条件下的性能差距呈简单线性对应关系,编码结构对语义信息的保留与利用能力在该阶段起到更为关键的作用。这表明,本文方法在低码率条件下的性能提升不仅依赖于提示词生成质量,更与所提出的编码框架设计密切相关。

3.8 运行复杂度与实现分析

在实验环境中,我们对T-ICM的编解码过程进行了时间统计。测试平台为NVIDIA RTX 4090 GPU,采用单样本(batch size=1)的推理配置,对完整的编码与解码流程进行测量。在上述设置下,本文方法的平均编码时间约为126 ms,平均解码时间约为79 ms。该统计结果反映了在当前实验配置下,T-ICM能够在合理的时间开销内完成一次完整的编解码过程。在显

存开销方面,T-ICM在上述配置下的峰值显存占用约为12 GB(包含特征提取、特征编解码及任务解码模块),能够在当前主流高性能GPU平台上稳定运行。

4 结束语

本文提出了一种面向机器视觉的文本提示引导的图像编码算法(T-ICM),针对传统图像编码方法在极低码率条件下性能退化严重,以及现有ICM框架在跨任务场景中泛化能力不足的问题,给出了新的解决思路。T-ICM通过语义-纹理解耦,将图像信息划分为紧凑的文本提示与通用特征两部分,并利用文本模态的高语义密度与强引导能力,在解码端实现对重建特征的自适应调整,从而在保证极低码率的同时保持了任务相关信息的完整性和可用性。实验结果表明,T-ICM在语义分割、实例分割、目标检测和图像分类等多个任务中均表现出显著的性能优势,相比传统编码器与基于深度学习的图像压缩方法,在低10倍甚至20倍码率条件下仍能保持竞争性能,特别是在0.017 bpp的极低码率下依然实现了接近原始输入的任务精度,展现出强大的实用价值。此外,T-ICM的分组特征编码模块(GFC)有效降低了模型参数量与训练复杂度,提示词生成器的模块化设计则赋予方法更高的灵活性与可扩展性,使其能够适配视觉分析模型、用户自定义输入以及多模态大语言模型等多种场景,充分体现了在开放任务和跨域应用下的鲁棒性。综上,T-ICM不仅在理论上拓展了ICM的研究边界,也在实践中展示了极低码率下的人机协同视觉处理潜力。未来工作将进一步探索提示生成的轻量化与自适应机制、跨模态应用的扩展,以及与现有视频编码标准的融合与硬件实现,为构建高效、智能和可部署的下一代机器视觉编码系统提供新的方向。

参考文献

[1] Bross B, Wang Yekui, Ye Yan, et al. Overview of the versatile video coding (VVC) standard and its applications[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(10): 3736-3764.

[2] Zhang Jiaqi, Jia Chuanmin, Lei Meng, et al. Recent development of AVS video coding standard: AVS3[C]//2019 Picture Coding Symposium. Piscataway: IEEE, 2019: 1-5.

[3] Ballé J, Laparra V, Simoncelli E P. End-to-end optimized image compression[C/OL]//Proceedings of the 5th International Conference on Learning Representations, ICLR, 2017, https://openreview.net/forum?id=rJxdQ3jeg.

[4] 董浩, 李劭辉, 阚诺文, 等. 基于深度压缩感知的联合信源信道编码方法研究[J]. 电子学报, 2025, 53(7): 2178-2192. Dong Hao, Li Shaohui, Kan Nuowen, et al. Research on joint source-channel coding method based on deep compressive sensing[J]. Acta Electronica Sinica, 2025, 53(7): 2178-2192. (in Chinese)

[5] Redondi A, Baroffio L, Bianchi L, et al. Compress-then-analyze versus analyze-then-compress: What is best in visual sensor networks[J]. IEEE Transactions on Mobile Computing, 2016, 15(12): 3000-3013.

[6] Huang Zhimeng, Jia Chuanmin, Wang Shanshe, et al. Visual analysis motivated rate-distortion model for image coding[C]//2021 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2021: 1-6.

[7] Bajić I V, Lin Weisi, Tian Yonghong. Collaborative intelligence: Challenges and opportunities[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2021: 8493-8497.

[8] Feng Ruoyu, Liu Jinming, Jin Xin, et al. Prompt-ICM: A unified framework towards image coding for machines with task-driven prompts[PP/OL]. V1.arXiv (2023-05-04)[2023-10-01]. https://doi.org/10.48550/arXiv.2305.02578.

[9] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]//The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Piscataway: IEEE, 2003: 1398-1402.

[10] 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. Piscataway: IEEE, 2018: 586-595.

[11] Stern M K, Johnson J H. Just noticeable difference[M]//Weiner I B, Craighead W E. The Corsini Encyclopedia of Psychology. 4th ed. Hoboken: John Wiley & Sons, 2010: 1-2.

[12] Fu C M, Alshina E, Alshin A, et al. Sample adaptive offset in the HEVC standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1755-1764.

[13] Tsai C Y, Chen C Y, Yamakage T, et al. Adaptive loop filtering for video coding[J]. IEEE Journal of Selected Topics in Signal Processing, 2013, 7(6): 934-945.

[14] Ballé J, Minnen D, Singh S, et al. Variational image compression with a scale hyperprior[PP/OL]. V1.arXiv (2018-12-01)[2025-10-26]. https://arxiv.org/abs/1802.01436.

[15] Cheng Zhengxue, Sun Heming, Takeuchi M, et al. Learned image compression with discretized Gaussian mixture likelihoods and attention modules[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 7936-7945.

[16] Lu Ming, Guo Peiyao, Shi Huiqing, et al. Transformer-based image compression[C]//2022 Data Compression Conference. SnowBird: IEEE, 2022: 469-469.

[17] Li H, Li S, Dai W, et al. Frequency-aware transformer for learned image compression[EB/OL]. (2024) [2025-10-26]. https://openreview.net/forum id=HKGQDDTuvZ.

[18] Theis L, Salimans T, Hoffman M D, et al. Lossy compression with Gaussian diffusion[PP/OL]. V2. arXiv (2022-12-31)[2025-08-26]. https://doi.org/10.48550/arXiv.2206.08889.

[19] Xia Yichong, Zhou Yimin, Wang Jinpeng, et al. DiffPC: Diffusion-based high perceptual fidelity image compression with semantic refinement[EB/OL]. (2025) [2025-10-26]. https://openreview.net/forum id=RL7PycCtAO.

[20] Mentzer F, Agustsson E, Tschannen M. M2T: Masking transformers twice for faster decoding[PP/OL]. V1.arXiv (2023-04-14)[2025-08-26]. https://doi.org/10.48550/arXiv.2304.07313.

[21] Cai Qi, Chen Zhifeng, Wu D O, et al. A novel video coding strategy in HEVC for object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(12): 4924-4937.

[22] Chen Zhibo, He Tianyu. Learning based facial image compression with semantic fidelity metric[J]. Neurocomputing, 2019, 338: 16-25.

[23] Le N, Zhang Honglei, Cricri F, et al. Image coding for machines: An end-to-end learned approach[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2021: 1590-1594.

[24] Patwa N, Ahuja N, Somayazulu S, et al. Semantic-preserving image compression[C]//2020 IEEE International Conference on Image Processing. Piscataway: IEEE, 2020: 1281-1285.

[25] Chamain L D, Racapé F, Bégaint J, et al. End-to-End optimized image compression for machines: A study[C]// 2021 Data Compression Conference. Piscataway: IEEE, 2021: 163-172.

[26] Le N, Zhang Honglei, Cricri F, et al. Learned image coding for machines: A content-adaptive approach[C]//2021 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2021: 1-6.

[27] Yan Ning, Gao Changsheng, Liu Dong, et al. SSSIC: Semantics-to-signal scalable image coding with learned structural representations[J]. IEEE Transactions on Image Processing, 2021, 30: 8939-8954.

[28] Feng Ruoyu, Jin Xin, Guo Zongyu, et al. Image coding for machines with omnipotent feature learning[C]//17th European Conference on Computer Vision-ECCV 2022. Heidelberg: Springer, 2022: 510-528.

[29] Chen Y H, Weng Y C, Kao C H, et al. TransTIC: Transferring transformer-based image compression from human perception to machine perception[C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 23240-23250.

[30] Xia Sifeng, Liang Kunchangtai, Yang Wenhan, et al. An emerging coding paradigm VCM: A scalable coding approach beyond feature and signal[C]//2020 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2020: 1-6.

[31] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 15979-15988.

[32] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[C/OL]//9th International Conference on Learning Representations, ICLR, https://iclr.cc/virtual/2021/poster/3013.

[33] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 779-788.

[34] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9630-9640.

[35] Yin Shukang, Fu Chaoyou, Zhao Sirui, et al. Woodpecker: Hallucination correction for multimodal large language models[J]. Science China Information Sciences, 2024, 67(12): 220105.

[36] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021: 8748-8763.

[37] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]// 2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 3992-4003.

[38] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2999-3007.

[39] Feng R, Qi Y, Liu J, et al. Diff-ICMH: Harmonizing machine and human vision in image compression with generative prior[EB/OL]. (2025) [2025-10-26]. https://openre-

view.net/forum id=ne3nYEcGsf.

[40] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[M]//13th European Conference on Computer Vision-ECCV 2014. Heidelberg: Springer, 2014: 740-755.

[41] Hong J, Fulton M, Sattar J. TrashCan: A semantically-segmented dataset towards visual detection of marine debris[PP/OL]. V1.arXiv (2020-07-16)[2025-08-26]. https://doi.org/10.48550/arXiv.2007.08097.

[42] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset[R]. California Institute of Technology, 2011.

[43] Khosla A, Jayadevaprakash N, Yao B, et al. Novel dataset for fine-grained image categorization[EB/OL]. (2011)[2025-10-26]. https://people.csail.mit.edu/khosla/papers/cub2011.pdf.

[44] Barman N, Martini M G, Reznik Y. Bjøntegaard delta (BD): A tutorial overview of the metric, evolution, challenges, and recommendations[PP/OL]. V1.arXiv (2024-01-08)[2025-08-26]. https://doi.org/10.48550/arXiv.2401.04039.

[45] Ghiasi G, Cui Yin, Srinivas A, et al. Simple copy-paste is a strong data augmentation method for instance segmentation[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 2917-2927.

[46] Minderer M, Gritsenko A, Stone A, et al. Simple open-vocabulary object detection[C]//Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 728-755.

[47] Liu Shilong, Zeng Zhaoyang, Ren Tianhe, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection[PP/OL]. V5.arXiv (2024-07-19)[2025-08-26]. https://doi.org/10.48550/arXiv.2303.05499.

作者简介

黄志勐　男,1997年6月出生于山东省济宁市。现为北京大学计算机学院博雅博士后。主要研究方向为智能编码、面向机器视觉的图像视频编码、多媒体技术和信号处理。

E-mail: zmhuang@pku.edu.cn

杨　帆　男,1992年10月出生于江西省瑞金市。现为北京大学艺术学院高级工程师。主要研究方向为多媒体与人工智能、计算艺术。

E-mail: fyang.eecs@pku.edu.cn

高　峰　男,1983年11月出生于北京市。现为北京大学艺术学院研究员、博士生导师、创意实验室主任。主要研究方向为计算机与艺术交叉学科,探索人类未来生活中人工智能技术在教育、艺术、健康等领域的应用。

E-mail: gaof@pku.edu.cn

马思伟　男,1979年2月出生于山东省聊城市。现为北京大学博雅特聘教授,北京大学计算机学院党委副书记、博士生导师,视频与视觉技术国家工程研究中心副主任。主要研究方向为视频编码与处理。中国电子学会会员编号:E190014267M。

E-mail: swma@pku.edu.cn