arXiv:2305.02586v2 [eess.IV] 3 Mar 2025

基于不规则分组解耦的语义结构化图像压缩

Ruoyu Feng[1*] Yixin Gao[1*] Xin Jin[2] Runsen Feng[1] Zhibo Chen [1,†]中国科学技术大学 [2]东部高等研究院

摘要

图像压缩技术通常侧重于压缩供人类使用的矩形图像，但这会导致为下游应用传输冗余内容。为克服这一局限，先前的一些工作提出对比特流进行语义结构化，通过选择性传输和重建来满足特定应用需求。然而，这些方法根据语义将输入图像划分为多个矩形区域，却未能避免区域间的信息交互，造成比特率浪费以及区域边界重建失真。本文提出基于定制组掩码将图像解耦为多个不规则形状的组，并对其独立压缩。我们的组掩码能够以更细的粒度描述图像，通过减少冗余内容的传输显著节省比特率。此外，为确保选择性重建的保真度，本文提出了组独立变换的概念，以维持不同组之间的独立性。我们通过提出的组独立 ***Swin*** 块（***GI Swin*** 块）对其进行实例化。实验结果表明，我们的框架能够以可忽略的开销构建比特流，并在视觉质量和智能任务支持方面均表现出卓越性能。代码可在 *https://github.com/Amygyx/GIT-SSIC*获取。

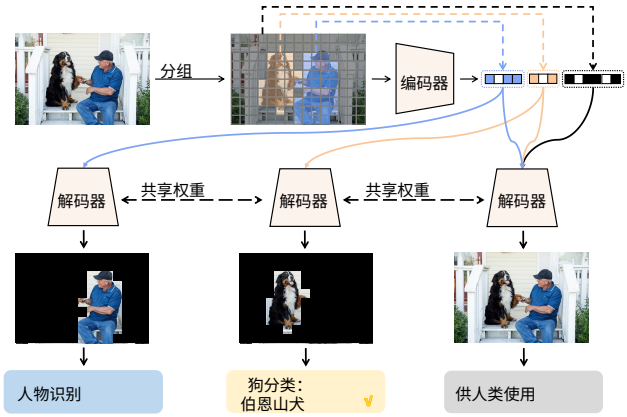


图 1. 输入图像根据不同语义解耦为多个组。随后，基于划分后的组对图像进行压缩，生成语义结构化比特流（SSB）。根据具体任务需求，SSB 可通过选择性传输比特流和部分重建来支持下游应用。

1. 引言

图像信号是现代社会各种应用的重要信息载体。图像压缩技术旨在将图像转换为紧凑的表示形式（即比特流），以节省传输和存储资源。有损图像压缩是最实用的技术之一，因为它能够在舍弃少量非必要信息的同时，恢复重要内容。

过去几十年来，传统图像压缩标准 [10,56,63,67,69]得到了广泛研究和应用。随着深度学习的快速发展，神经图像编解码器 [7, 8, 16, 28, 36, 42, 43, 52——54, 77]也迅速演进，并取得了令人瞩目的成果。与此同时，越来越多的媒体内容倾向于由机器视觉算法处理，例如识别[17, 27, 30, 34, 49], 检测 [45, 46, 57——59], 和分割 [5, 9, 11, 12, 29, 48, 50, 71, 75]。然而，大多数压缩方法主要针对供人类观看的规则矩形图像而开发，并未考虑其在下游任务或人机交互场景中的效率与功能。

近年来，面向机器的图像编码（ICM）领域应运而生，旨在开发一种高效且具备分析能力的联合框架，以支持智能分析。针对特定任务损失与压缩率之间的权衡进行端到端优化，是一种很有前景的方法 [4, 13, 14, 18, 19, 40,

* 前两位作者贡献相同。
† 通讯作者。

41, 51, 62], 但其缺乏对大规模多样化应用的泛化能力。为克服这一局限，Feng等人提出在熵约束下，压缩通过自监督学习获得的通用且紧凑的特征，以支持下游任务。然而，这需要以所提出的特征作为输入重新训练任务模型，使整体性能严重依赖于特征提取的有效性。上述方法专门针对机器视觉中的压缩场景设计，未考虑需要人工参与的情况。语义结构化图像压缩（SSIC） [64]提出通过使用预先准备的目标检测工具箱，分别压缩检测到的对象的矩形区域，从而生成语义结构化比特流（SSB）。尽管SSB能够通过比特流的部分传输和重建，高效地支持智能任务和人机交互，但其基于矩形区域的划分方法可能会遇到对象重叠的问题。SSIC通过用更大的矩形区域替换重叠对象来解决这一问题，但这可能造成比特率浪费。此外，SSIC直接在潜在域中压缩相应的潜变量，以生成每个对象的比特流，而未考虑变换过程中特征之间的交互与依赖。这可能导致部分重建场景中的组边界模糊和失真，进而影响重建质量。

超越基于矩形的划分，本文提出根据定制的组掩码，将图像解耦为多个不规则形状的组。随后，通过独立压缩这些组生成 SSB，并可通过选择性传输和重建支持各种需求。值得注意的是，组掩码的生成在组的形状、预分析方式和划分标准方面具有高度灵活性，从而能够根据不同的应用场景和需求进行定制。此外，为避免部分重建场景中潜在的质量下降，我们提出了组独立变换的概念，确保潜在表示中各组之间的独立性，因此选择性重建的质量不会受到其他组缺失的影响。更具体地说，我们通过精心设计组独立 Swin 块（GI Swin 块）来实现该概念。GI Swin 块是针对我们的场景和需求定制的 Swin Transformer[49] 扩展。GI Swin 块利用 Swin Transformer [49, 77], 的层次化建模能力，在保证组独立性的前提下实现高编码效率。通过结合基于组掩码的解耦与组独立 Swin 块，我们提出的方法能够高效支持包括仅通过生成一个比特流，即可支持人机交互和机器视觉任务。

我们方法的主要贡献总结如下：

- 我们提出将图像解耦为多个不规则形状的组，以实现比特流的结构化。我们的组掩码能够以比典型矩形更精细的粒度描述空间划分，并通过减少冗余内容传输来节省比特率。
- 我们提出了组独立变换，并通过精心设计组独立 Swin 块（GI Swin 块）将其实例化，该模块既保持了强大的变换能力，又确保了潜在表示中各组之间的独立性。
- 实验结果表明，我们提出的模型具有最先进的压缩能力和卓越的下游任务性能，是一种兼具高压缩效率与功能性的编解码器。

2. 相关工作

2.1. 图像压缩

传统图像压缩。 经过数十年的发展，JPEG [67], JPEG2000 [56],HEVC [63], 和 VVC 帧内编码 [10], 等传统图像压缩标准已在实践中得到广泛应用。这些标准依赖于变换编码 [24],, 将有损图像压缩任务分解为三个部分：变换、量化和熵编码。这些标准的每个模块都由人工设计，包含多种模式，并通过率失真优化来确定最佳模式。然而，传统编解码器完全由人工设计的结构限制了其灵活性和可扩展性，使其难以支持多种目标，例如多尺度结构相似性和分类准确率，因为它们无法以端到端的方式进行优化。

学习式图像压缩。 近年来，基于非线性变换编码的学习式图像压缩方法 [6]取得了快速进展。该领域的早期研究主要通过开发微分量化和码率估计技术 [1, 7, 65] 来实现端到端训练。随后，大量研究设计了强大的神经网络模块，以增强变换 [16, 77], 、量化 [26, 72], 和熵模型 [8, 28, 54, 55]。因此，一些神经图像编解码器 [16, 25, 28, 54, 55] 能够达到与 HEVC [63] 和 VVC [10]等传统编码标准相当甚至更优的性能。此外，一些

研究 [2,53,60] 引入感知损失，以有效提升重建质量。然而，现有的图像压缩方法侧重于压缩整幅图像，未考虑对比特流进行选择性传输和重建，以实现任意重建；因此，在面向具有不同需求的下游应用时，会造成大量带宽浪费。

2.2. 面向机器的图像编码

面向机器的图像编码旨在压缩源图像，以服务于识别[17, 27, 30, 34, 49], 检测 [45, 46, 57–59], 和分割等下游任务 [5, 9, 11, 12, 29, 48, 50, 71, 75]。以端到端的方式联合优化任务损失和比特率 [3, 32, 40, 44, 68]，是一种自然且前景广阔的方法。另一种思路是压缩与下游任务相对应的特征 [4,13,14,18,19,51,62]。然而，这些方法往往偏向特定任务，对于广泛的智能应用可能缺乏泛化能力。为解决泛化能力不足的问题，有研究提出了一种基于表征学习的新方法 [21]，将对比学习与熵约束相结合，以学习通用且紧凑的特征。学习到的特征被用来取代原始图像，作为压缩和传输的新源，从而显著提升了各种智能任务的编码效率。然而，这种方法需要使用所提出的特征作为输入重新训练任务模型，而且其整体性能很大程度上取决于特征提取的效果。本文提出了一种统一的压缩框架，无需进行面向任务的联合训练或额外的特征学习，即可支持机器视觉。此外，该框架还可根据需求进行完整或选择性重建，以满足人类感知需求。

2.3. 面向对象的图像压缩

面向对象的图像压缩最早由 MPEG-4 Visual 提出，通过压缩任意形状的视觉对象平面（VOP），旨在为图像压缩技术提供基于内容的交互性和可扩展性。然而，由于难以标注目标复杂而精细的形状，且缺乏实际的下游应用，MPEG-4 Visual 并未得到广泛应用。近年来，随着深度学习的快速发展，面向下游应用的图像压缩实际场景迅速增多。Sun等人提出了语义结构化图像压缩（SSIC），通过将图像解耦为多个对象，并使用神经图像编解码器独立压缩相应的潜变量，生成语义结构化比特流。然而，SSIC 采用的基于矩形的划分难以泛化到不规则对象，也缺乏定制灵活性，此外，直接在潜在空间中选择对象的表示会忽略编码变换过程中各元素之间的相互作用，导致对象在部分重建时出现明显的边缘失真。本文提出根据组掩码将图像解耦为灵活且可定制的不规则组，从而实现语义结构化图像压缩。此外，我们提出一种组间独立变换，在编码和解码过程中均可执行，从而能够根据不同需求高效地部分或完整重建图像。

3. 方法

3.1. 概述

所提出的方法是语义结构化图像压缩的一种高效且灵活的版本 [64]。通过由块级组掩码引导传输和重建部分空间区域，该方法能够满足多种应用需求。我们的模型以组掩码作为引导，确保变换操作之间的交互仅发生在同一组内，从而在冗余去除过程中实现组间独立。组掩码基于预分析生成，例如目标检测、语义分割和显著性检测，为比特流的结构化提供了高度的灵活性和可定制性，并被视为旁信息。随后，分别对每个组的变量执行熵编码过程，从而生成按语义结构化的比特流。根据解码端的需求，比特流可以部分或全部传输；接收端随后对比特流执行熵解码，并根据组掩码和组索引重新组织潜变量的空间排列。

整体网络结构如图所示。2，其中包含组掩码和 GI Swin 块，分别用于引导和实现组独立变换。第3.2节和第3.3节分别介绍组掩码和组独立变换。网络架构的详细实现见第3.4节。

3.2. 组掩码

压缩前，根据目标检测、实例分割和显著性检测等预分析技术的结果生成组掩码。图 3 展示了这一过程的示例。更具体地说，对于高度为 x 、宽度为 H 的输入图像 W，组掩码的空间分辨率与输入图像相同，由 $\frac{H}{B} \times \frac{W}{B}$ 个块组成，其中 B 为每个块的边长。并且

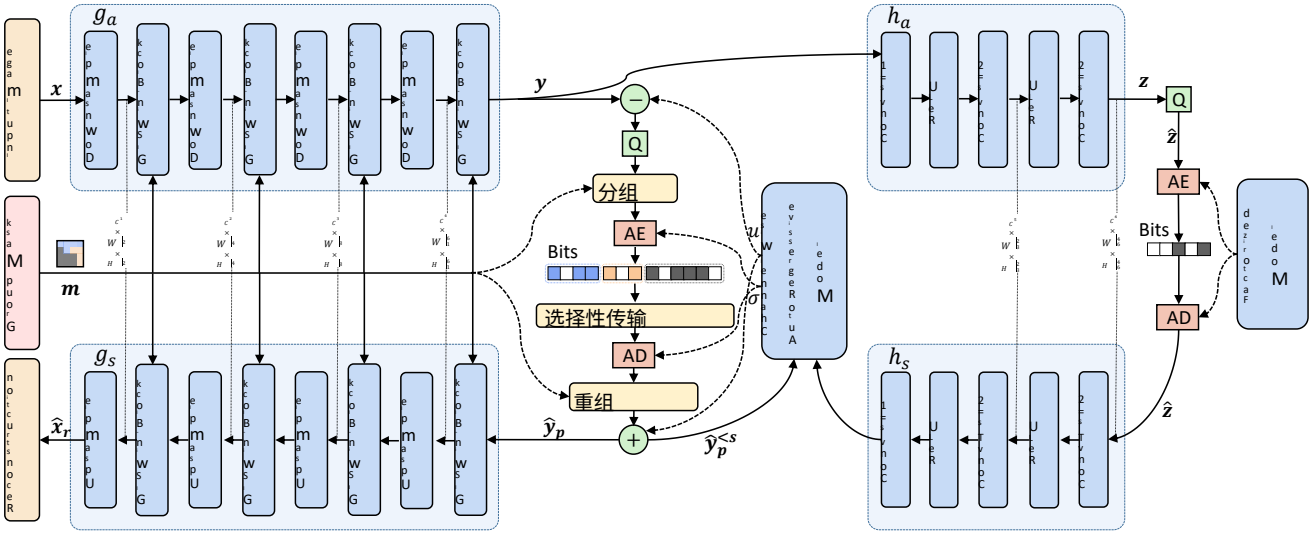


图 2. 我们提出的模型与通道自回归模型（ChARM）的网络架构。ConvT 表示转置卷积。AE 和 AD 分别表示算术编码和算术解码。在 Ours-Hyper模型中，我们移除 ChARM 组件，改为由超解码器 h_s直接输出 μ 和 σ 。

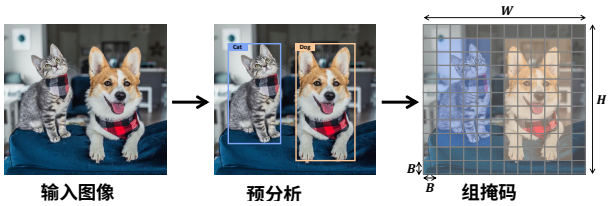


图 3. 组掩码生成示例。首先通过目标检测对输入图像进行预分析，然后根据预分析结果生成组掩码。

它表示相应组的索引。

我们提出的组掩码可以清晰地划分经过单次下采样变换后潜变量空间中的元素。此外，在进行压缩之前，将组掩码下采样 B 倍可以大幅降低开销比特率。块大小与开销比特率呈反比关系。此外，如图4所示，组掩码的生成方式可以灵活定制，包括预分析的方式和块分配的标准等。

3.3. 组独立变换

基于变换编码的有损图像压缩可分为三个模块化组件：变换、量化和熵编码。量化和熵编码不会影响语义结构化比特流（SSB）的生成与使用，而传统变换作用于整幅图像以去除空间冗余，必然会在压缩过程中产生组间依赖。在选择性传输和重建的情况下，由对所有组中的部分组进行重新组织所产生的不完整表示，可能会因缺少组间依赖而导致重建不准确。因此，我们提出了用于生成语义结构化比特流的组独立变换这一概念。其基本思想是将变换限制为仅在每个组内进行。

一种直接且自然的方法是使用变换器[66] ，并采用与组掩码相对应的定制注意力图。然而，图像压缩与变换器擅长的高层理解任务存在显著差异。大量复杂的长程依赖关系可能阻碍收敛，并对最终性能产生不利影响。为了在保持组独立属性的同时实现高编码效率，并受到 Swin-Transformer 强大的层次化表示建模能力的启发[49, 77], ，我们将 Swin-Transformer 的 Swin 块扩展为所提出的组独立 Swin 块（GI Swin 块）。GI Swin 块是变换的核心组件。具体而言，如图5a所示，通过将窗口划分和组划分合并为组独立窗口划分，使自注意力能够在划分后的局部区域内计算，从而在保持组独立属性的同时实现高编码效率。此外，如图5b所示，通过以类似方式合并移位窗口划分和组划分，引入了跨窗口连接。

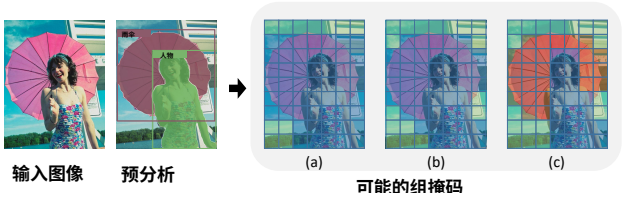


图 4. 组掩码自定义灵活性的示意图：(a) 在边界框的引导下，将重叠对象作为一个组。(b) 在实例掩码的引导下，将重叠对象作为一个组。(c) 在实例掩码的引导下，将重叠对象作为不同的组。

3.4. 网络架构与流程管线

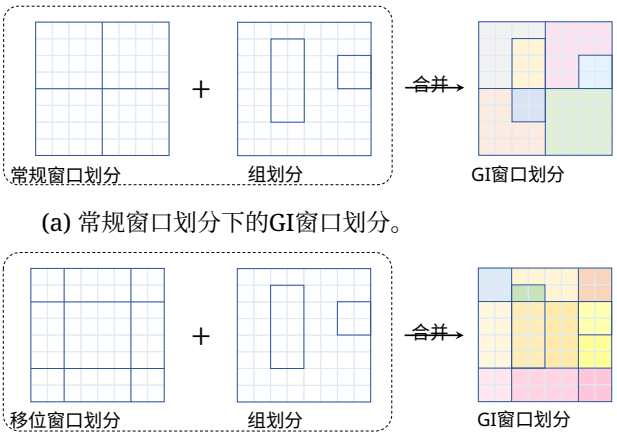
图2展示了所提出的架构。通过结合上采样/下采样和组独立 Swin 块（GI Swin 块），以交错方式执行主要变换。上采样包含一个像素重排操作和一个具有1×1卷积核大小的卷积层，而下采样则包含一个像素逆重排操作和一个具有 1×1 卷积核大小的卷积层。在熵编码方面，我们采用了均值与尺度（M&S）超先验模型 [54]，并结合强大的通道自回归模型（ChARM） [55]来预测潜变量的概率分布。

具体而言，给定源图像 $\boldsymbol{x}$，编码器 g_a 将其转换为以给定的组掩码 $\boldsymbol{m} = \sum_{i=1}^{N} \boldsymbol{m}_i$为条件的潜在表示 $\boldsymbol{y}$，其中 $i \in \{1, 2, 3, \ldots, N\}$表示相应组在全部 N组中的索引。在每个 $\boldsymbol{m}_i$中，与 i_{th} 组对应位置的值被设为 i，而其他位置的所有值均设为 0。类似地，我们可以从组的角度定义$\boldsymbol{x} = \sum_{i=1}^{N} \boldsymbol{x}_i$ 和$\boldsymbol{y} = \sum_{i=1}^{N} \boldsymbol{y}_i$，且变换仅在每个组内进行，表示为

$$\boldsymbol{y}_i = g_a\left(\boldsymbol{x}_i | \boldsymbol{m}_i\right). \tag{1}$$

得益于组独立变换，我们可以轻松地在一次前向传播中实现所有组的变换，因此，该过程也可以表示为 $\boldsymbol{y} \quad = g_a(\boldsymbol{x}|\boldsymbol{m})$。

随后，将潜变量量化为离散表示$\hat{}\boldsymbol{y}$ ，并通过熵模型根据预测的概率分布有选择地传输至解码器端。具体而言，超编码器 $\boldsymbol{z} = h_a(\boldsymbol{y})$ 提取辅助信息 $\boldsymbol{z}$ 。量化后的超潜变量$\hat{}\boldsymbol{z} = Q(\boldsymbol{z})$通过学习得到的因子化先验进行建模和熵编码。在 ChARM 中，潜变量$\hat{}\boldsymbol{y}$ 沿通道维度切分为 S 个切片（我们选择 S=10），并根据先前的切片$\hat{}\boldsymbol{y}_{<s}$对每个切片$\hat{}\boldsymbol{y}_s$ 进行熵编码。$\hat{}\boldsymbol{y}$ 通过与单位均匀噪声卷积的条件高斯分布进行建模。



(a) 常规窗口划分下的GI窗口划分。

(b) 移位窗口划分下的 GI窗口划分。

图 5. 常规窗口划分和移位窗口划分下 GI Swin-Block 的组无关窗口划分。在每个局部区域内进行自注意力。

对量化后的潜变量进行选择性熵解码后，我们得到$\hat{}\boldsymbol{y}_p = \sum_{j \in \mathbf{P}} \hat{}\boldsymbol{y}_j$，其中$\mathbf{P} \subseteq \{1, 2, \ldots, N\}$ 是所需组的汇集。重建图像$\hat{}\boldsymbol{x}_p = \sum_{j \in \mathbf{P}} \hat{}\boldsymbol{x}_j$ 将按如下方式计算

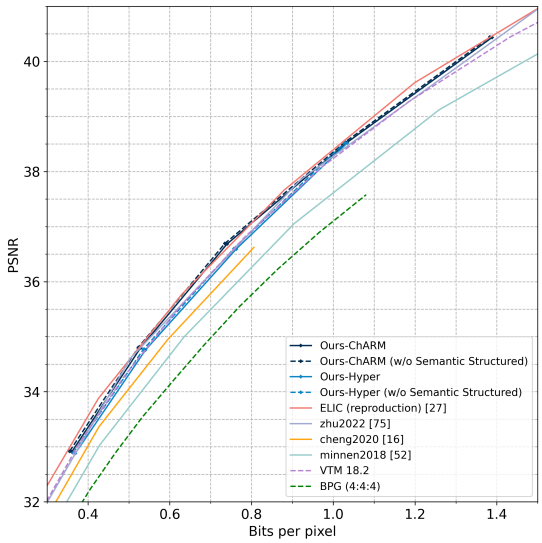
$$\hat{\boldsymbol{x}}_j = g_s\left(\hat{\boldsymbol{y}}_j | \boldsymbol{m}_j\right). \tag{2}$$

与方程（1）类似，方程（2）也可以写为$\hat{}\boldsymbol{x} = g_s(\hat{}\boldsymbol{y}|\boldsymbol{m})$。请注意，舍入操作是不可微分的，因此我们在熵模型的学习过程中采用均匀噪声，并使用舍入后的结果作为解码器 g_s 的输入，如 [55]所示。

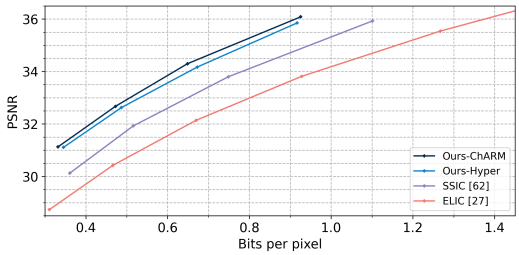
4. 实验

4.1. 实验设置

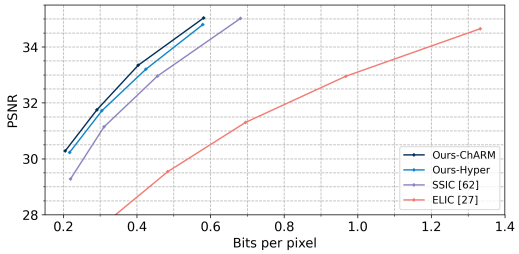
图像编解码器。 我们在 Flickr 2K 数据集上训练高比特率模型（>=0.6bpp），其设置与 [76] 相同；并在 COCO 2017 训练集上训练低比特率模型（$<$ 0.6bpp） [47]。Ours-Hyper 训练 320 万次迭代，而 Ours-ChARM 模型则通过从更高比特率的超先验模型初始化权重，再训练 120 万次迭代。每个批次包含从训练数据集中随机裁剪的 8 个 256×256 图像块。学习率设为 $5e-5$，并在经过 $2.8M$ 次迭代后衰减至原来的十分之一。训练损失$L = R + \beta D$ 是率失真权衡的加权组合，由拉格朗日乘子β进行平衡。我们采用 RGB 色彩空间中的均方误差（MSE）作为失真度量 D。通过应用不同的 $\beta\,\beta \in \{512, 1024, 2048, 4096\}$，实现图像编解码器的比特率范围。在训练阶段，我们随机生成组掩码，用于训练组独立变换；具体操作



(a) Kodak 上的整幅图像重建。



(b) COCO 上所有类别对象的重建。



(c) COCO 上人体类别对象的重建。

图 6. 整幅图像和部分重建场景下压缩效率的率失真对比。部分重建的峰值信噪比仅在预检测的边界框上计算。

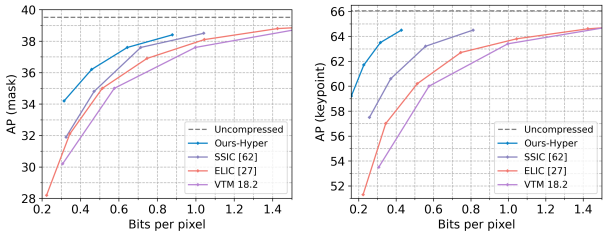


图 7. COCO 上实例分割（左）和姿态估计（右）的性能对比。

见补充材料。

评估数据集与协议。 对于整体图像重建质量，我们使用广泛采用的 Kodak 数据集 [39]评估模型的编码效率。我们使用从 COCO 2017 验证集中随机选取的 40 张图像和 20 张图像，分别评估所有类别对象及人体对象的选择性重建质量。感兴趣区域的峰值信噪比（PSNR）作为衡量客观质量的指标。此外，每像素比特数（bpp）通过将传输比特流的总比特数除以感兴趣区域的像素数计算得出。为验证所提出方法对下游任务的支持效果，我们使用 COCO 2017 验证集，在实例分割和姿态估计任务上开展实验。 [47]。该数据集广泛用于密集预测任务，包含 118K 张训练图像、5K 张验证图像以及 20K 张 test-dev 图像。我们分别使用 Mask R-CNN（X101-FPN 主干网络）和 Keypoint R-CNN（X101-FPN 主干网络），通过 Detectron2 工具箱实现实例分割和人体关键点检测性能评估 [70]。

4.2. 定量结果

在进行比较时，我们选择了近期的神经图像编解码器 [16,28,54,77], 、强大的传统图像编解码器 [10, 63]，以及支持语义结构化比特流的功能性编解码器 [64]。其中，ELIC [28] 经复现后，性能与其报告结果相近。需要着重指出的是，编码器并不具备解码器所需感兴趣区域的先验知识。默认情况下，我们采用目标检测作为预分析手段，这是最基础的视觉分析工具之一。语义结构化比特流根据 Mask R-CNN（X101-FPN 主干网络）检测器检测到的边界框生成，并依据相关语义的要求选择性地解码和重建为图像，以供评估。在所有定量结果实验中，组掩码的块大小 B 设为 32，以平衡性能与开销比特率之间的权衡；重叠的组则取其并集作为一个组。补充材料中还展示了与其他面向机器的图像编码方法进行比较的更多实验。

重建质量。对于整幅图像重建，图。6a 表明，我们提出的模型实现了业界领先的率失真性能，在所有比特率下的峰值信噪比均超过 VTM 18.2。值得注意的是，按照语义对比特流进行结构化只会引入可忽略的影响。因此，我们的模型在同时实现高编码效率和功能性方面具有巨大潜力。对于特定感兴趣场景中的部分图像重建，如图。6b 和图。6c所示，与其他编解码器相比，我们的模型取得了显著改进。具体而言，语义-

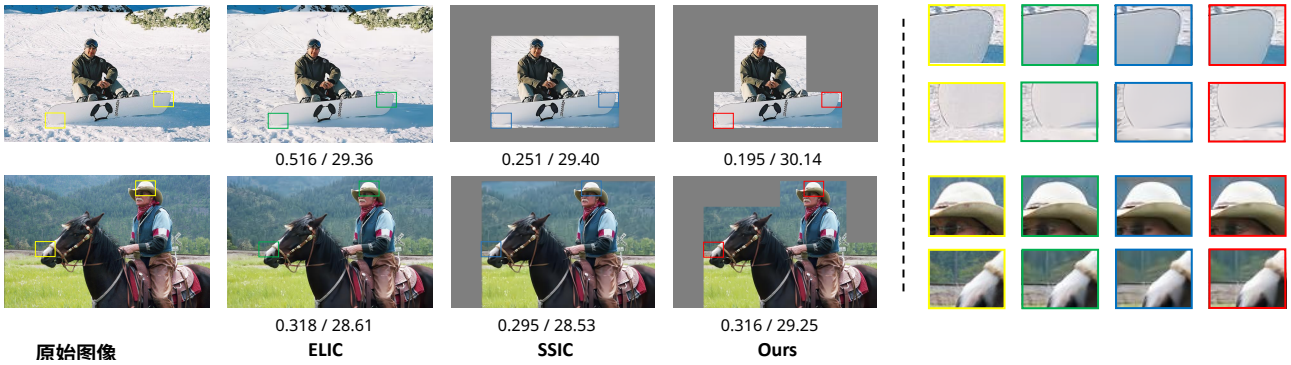


图 8. 在以前景对象为感兴趣区域的场景中，COCO 2017 数据集上供人类观看的重建图像可视化结果。图像下方的数字对应每像素比特数和峰值信噪比。

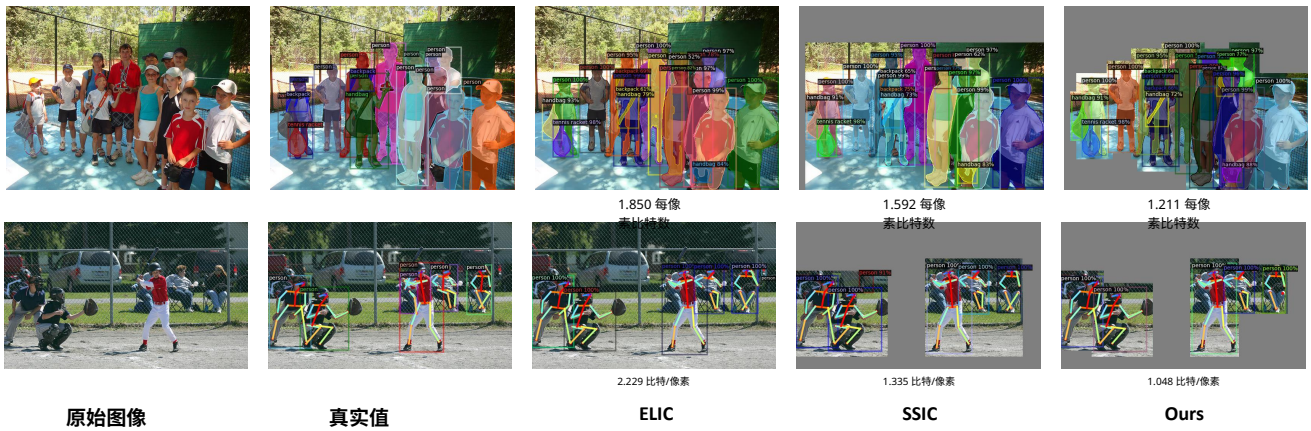


图 9. COCO 2017 数据集中用于下游任务的重建图像可视化结果。第一行为实例分割，第二行为姿态估计。

结构化比特流使我们的模型和 SSIC [64]能够避免传输和解码与整幅图像对应的比特流，而 ELIC 则需要这样做 [28]。此外，与 SSIC 相比，我们模型灵活的块级组掩码和组独立变换在空间上将潜变量与语义层级解耦，从而生成更清晰、更逼真的边界 [64]。此外，对于区域重叠的情况，我们的模型可以将这些区域归入一个不规则组，而非用更大的边界框替代它们，从而显著节省比特率；后者会引入更多失真。

下游任务支持。图。7展示了我们在实例分割和姿态估计方面的卓越性能。在实例分割任务中，与其他方法相比，我们的模型在低比特率（<0.6bpp）下取得了显著提升的性能，这可归因于我们的方法能够忠实地重建 RoI 边界。值得注意的是，我们的方法在姿态估计方面表现出色，这是因为我们的模型能够保留精细细节和边界，而这些对于准确定位人体关键点至关重要。此外，人体对象在图像中占比较少，且通常呈稀疏分布。与 SSIC 相比，我们的

该方法通过避免传输对应于包含所有重叠对象的矩形区域的潜变量，可显著降低比特率。

4.3. 定性结果

面向人类消费的重建。如图8所示，当重建图像主要面向具有特定语义兴趣的人类消费时，针对常规矩形图像压缩而设计的编解码器所生成的比特流必须完整传输并用于完整重建，而不考虑图像的需求和内容。尽管 SSIC[64]通过基于边界框的划分生成 SSB，部分解决了这一问题，但其灵活性不足，并且在将重叠对象合并到更大的边界框中时，可能引入额外的不相关干扰。此外，SSIC 通过在基于卷积神经网络的变换之后，直接划分潜变量空间中的相应位置来生成比特流，这可能导致选择性重建过程中依赖关系丢失，从而造成区域边界模糊和失真。借助基于组掩码的分区和组独立变换，我们的方法能够更高效地解耦图像以生成 SSB，同时确保选择性重建过程中不会引入失真或模糊，



图 10. 从语义结构化比特流中进行的选择性重建。比特流可基于不同的预分析方法生成。



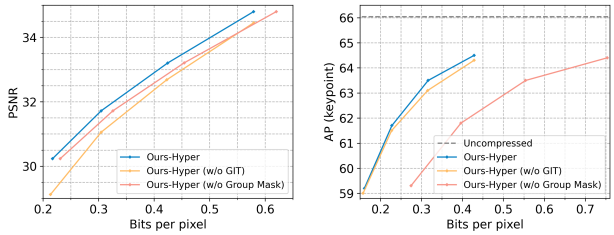
图 11. 基于我们提出的方法进行语义感知加密的示意图。

从而显著节省比特率并带来更佳的视觉体验。

面向智能任务的重建。 将重建图像用于下游智能任务时，基于相对语义的先验知识，有选择地传输并解码语义结构化比特流，可以显著节省比特率。如图9所示，传统编解码器旨在压缩矩形图像，因此需要传输并解码完整的比特流。随后，下游模型会对完全重建的图像进行智能分析，造成比特率浪费。与面向人类观看的图像重建类似，SSIC 也可以通过选择性传输节省比特率。然而，其基于边界框的分区方式可能无法最优地节省比特率，而且由此产生的模糊、失真的区域边界以及无关内容，可能进一步妨碍下游智能分析。（例如，在图 9中将汽车误识别为人物。）我们的方法能够更高效地生成 SSB，并以兼顾准确性和编码效率的方式支持具有特定需求的各类下游任务。

可定制性与灵活性。可以根据不同的语义划分标准，灵活定制语义结构化比特流。例如，如图10所示，语义结构化比特流的生成和使用既可以基于目标检测，也可以基于实例分割，具体取决于实际需求。需要指出的是，预分析方法并不限于图10所示的目标检测和实例分割，还可以采用显著性检测 [23, 31, 74], 语义分割 [5, 50, 71, 73, 75], 全景分割 [15, 38], 等，甚至可以采用人工标注。

语义感知加密。 如图所示11，我们提出的方法可应用于比特流加-



（a）感兴趣区域重建。 （b）支持智能任务。
图 12. 消融研究。

加密，从而能够根据用户的安全级别，利用语义先验进行选择性加密，甚至分层加密。在某些情况下，对 SSB 进行选择性加密，可以在尽量降低对视觉质量和下游分析影响的同时，安全传输和存储敏感信息。实现细节请参阅补充材料。

5. 消融研究

为了展示所提出的组掩码和组独立变换（GIT）在面向感兴趣区域的选择性重建和下游智能分析中的有效性，我们在不同码率下进行了消融研究。具体而言，我们将人物作为感兴趣区域进行感兴趣区域重建；将所有前景对象作为感兴趣区域的结果则见补充材料。此外，我们还针对 COCO 2017 上的姿态估计进行了支持智能任务的消融研究，实例分割的结果也收录于补充材料中。

组掩码。图14 显示，与基于朴素边界框的分区相比，基于组掩码的分区可显著节省比特率。同时，感兴趣区域重建的视觉质量和下游任务性能都不会受到影响。

组独立变换。如图14a所示，组独立变换对感兴趣区域重建的视觉质量至关重要。如图14b所示，如果不进行组独立变换，模糊且失真的区域边界会影响下游任务性能。

6. 结论

在本文中，我们提出了一种基于组掩码生成具有强大功能的语义结构化比特流的方法。组掩码具有高度的灵活性和可定制性，同时保持了轻量级开销。此外，我们首次提出了组独立变换的概念，并通过设计组独立 Swin 块（GI-Swin 块）将其实例化，以确保不同组之间的独立性，从而实现感兴趣区域的良好重建。特别是，我们提出的方法优于 VTM-18.2，并

且在提供强大功能的同时，实现了与先进神经图像编解码器相当的编码效率。实验结果证明了我们提出的方法在多种应用中的有效性，包括供人类使用的感兴趣区域、下游任务支持以及语义感知加密。

参考文献

[1] Eirikur Agustsson、Fabian Mentzer、Michael Tschannen、Lukas Cavigelli、Radu Timofte、Luca Benini 和 Luc V Gool。用于端到端学习可压缩表征的软到硬向量量化。NeurIPS，30，2017。2[2] Eirikur Agustsson、Michael Tschannen、Fabian Mentzer、Radu Timofte 和 Luc Van Gool。用于极致学习式图像压缩的生成对抗网络。载于 ICCV，第 221–231 页，2019。3[3] Mohammad Akbari、Jie Liang 和 Jingning Han。Dsslic：基于深度语义分割的分层图像压缩。载于 ICASSP 2019—2019 IEEE 国际声学、语音与信号处理会议（ICASSP），第 2042–2046 页。IEEE，2019。3[4] Artem Babenko、Anton Slesarev、Alexandr Chigorin 和 Victor Lempitsky。用于图像检索的神经编码。载于 ECCV，第 584–599 页。Springer，2014。1，3[5] Vijay Badrinarayanan、Alex Kendall 和 Roberto Cipolla。Segnet：用于图像分割的深度卷积编码器—解码器架构。IEEE模式分析与机器智能汇刊，39(12)：2481–2495，2017。1，3，8[6] Johannes Ball´e、Philip A Chou、David Minnen、Saurabh Singh、Nick Johnston、Eirikur Agustsson、S ung Jin Hwang 和 George Toderici。非线性变换编码。IEEE信号处理专题期刊，15(2)：339–353，2020。2[7] Johannes Ball´e、Valero Laparra 和 Eero P Simoncelli。端到端优化图像压缩。载于 ICLR，2017。1，2[8] Johannes Ball´e、David Minnen、Saurabh Singh、Sung Jin Hwang 和 Nick Johnston。采用尺度超先验的变分图像压缩。载于 ICLR，2018。1，2[9] Daniel Bolya、Chong Zhou、Fanyi Xiao 和 Yong Jae Lee。Yolact：实时实例分割。载于 ICCV，第 9157–9166 页，2019。1，3[10] Benjamin Bross、Ye-Kui Wang、Yan Ye、Shan Liu、Jianle Chen、Gary J Sullivan 和 Jens-Rainer Ohm。通用视频编码（VVC）标准及其应用概述。TCSVT，2021。1，2，6[11] Liang-Chieh Chen、George Papandreou、Iasonas Kokkinos、Kevin Murphy 和 Alan L Yuille。DeepLab：采用深度卷积网络、空洞卷积和全连接条件随机场的语义图像分割。IEEE模式分析与机器智能汇刊，40(4)：834–848，2017。1，3[12] Liang-Chieh Chen、Yukun Zhu、George Papandreou、Florian Schroff 和 Hartwig Adam。采用空洞可分离卷积的语义图像分割编码器—解码器。载于 ECCV，第 801–818 页，2018。1，3[13] Zhuo Chen、Kui Fan、Shiqi Wang、Lingyu Duan、Weisi Lin 和 Alex Chichung Kot。迈向智能感知：在-

中间深度特征压缩。IEEE图像处理汇刊，29：2243，2019。2230 1, 3 – . [14] Zhuo Chen、Kui Fan、Shiqi Wang、Ling-Yu Duan、Weisi Lin 和 Alex Kot。中间深度学习特征的有损压缩与评估。载于 ACM多媒体大会，第 2414–2422 页，2019。1, 3 – . [15] Bowen Cheng、Maxwell D Collins、Yukun Zhu、Ting Liu、Thomas S Huang、Hartwig Adam 和 Liang-Chieh Chen。Panoptic-DeepLab：一种简单、强大且快速的自底向上全景分割基线。载于 CVPR，第 12475–12485 页，2020。8[16] Zhengxue Cheng、Heming Sun、Masaru Takeuchi 和 Jiro Katto。基于离散化高斯混合似然和注意力模块的学习式图像压缩。载于 CVPR，第 7939–7948 页，2020。1, 2, 6[17] Alexey Dosovitskiy 、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly 等。图像胜过千言万语：用于大规模图像识别的 Transformer。I CLR，2020。1, 3[18] Ling-Yu Duan、Vijay Chandrasekhar、Jie Chen、Jie Lin、Zhe Wang、Tiejun Huang、Bernd Girod 和 Wen Gao。MPEG-CDVS标准概述。IEEE图像处理汇刊，25(1)：179–194，2015。1, 3[19] Ling-Yu Duan、Yihang Lou、Yan Bai、Tiejun Huang、Wen Gao、Vijay Chandrasekhar、Jie Lin、Shiqi Wang 和 Alex Chichung Kot。用于视频分析的紧凑描述符：新兴的 MPEG 标准。IEEE《多媒体》，26(2)：44–54，2018。1, 3[20] Touradj Ebrahimi 和 Caspar Horne。MPEG-4 自然视频编码概述。信号处理：图像通信，15(4-5)：365–385，2000。3[21] Ruoyu Feng、Xin Jin、Zongyu Guo、Runsen Feng、Yixin Gao、Tianyu He、Zhizheng Zhang、Simeng Sun 和 Zhibo Chen。通过全能特征学习实现面向机器的图像编码。载于 ECCV，第 510–528 页。Springer，2022。2, 3, 12[22] Ronald Aylmer Fisher、Frank Yates 等。生物学、农业和医学研究统计表，由 ra fisher 和 f. yates 编。爱丁堡：Oliver and Boyd，1963。12[23] Stas Goferman、Lihi Zelnik-Manor 和 Ayellet Tal。上下文感知显著性检测。IEEE模式分析与机器智能汇刊，34(10)：1915–1926，2011。8[24] Vivek K Goyal。变换编码的理论基础。IEEE《信号处理杂志》，18(5)：9–21，2001。2[25] Zongyu Guo、Zhizheng Zhang、Runsen Feng 和 Zhibo Chen。用于学习式图像压缩的因果上下文预测。TCSVT，32(4)：2329–2341，2021。2[26] Zongyu Guo、Zhizheng Zhang、Runsen Feng 和 Zhibo Chen。先软后硬：重新思考神经图像压缩中的量化。载于 ICML，第 3920–3929 页。PMLR，2021。2[27] Kai Han、An Xiao、Enhua Wu、Jianyuan Guo、Chunjing Xu 和 Yunhe Wang。Transformer中的Transformer。NeurIPS，34：15908–15919，2021。1, 3[28] Dailan He、Ziming Yang、Weikun Peng、Rui Ma、Hongwei Qin 和 Yan Wang。ELIC：采用不均匀分组空域—通道上下文自适应编码的高效学习式图像压缩。载于 CVPR，第 5718–5727 页，2022。1, 2, 6, 7

[29] Kaiming He、Georgia Gkioxari、Piotr Doll´ar 和 Ross Girshick。Mask R-CNN。载于 ICCV，第 2961–2969 页，2017。1, 3[30] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。载于 CVPR，第 770–778 页，2016。1, 3[31] Xiaodi Hou 和 Liqing Zhang。显著性检测：一种谱残差方法。载于 CVPR，第 1–8 页。IEEE，2007。8[32] Yueyu Hu、Shuai Yang、Wenhan Yang、Ling-Yu Duan 和 Jiaying Liu。面向人类与机器视觉的编码：一种可扩展的图像编码方法。载于 2020 IEEE 多媒体与博览会国际会议（ICME），第 1–6 页。IEEE，2020。3[33] Zhimeng Huang、Chuanmin Jia、Shanshe Wang 和 Siwei Ma。以视觉分析为导向的图像编码率失真模型。载于 2021 IEEE 多媒体与博览会国际会议（ICME），第 1–6 页。IEEE，2021。12[34] Menglin Jia、Luming Tang、Bor-Chun Chen、Claire Cardie、Serge Belongie、Bharath Hariharan 和 Ser-Nam Lim。视觉提示微调。ECCV，2022。1, 3[35] Xin Jin、Ruoyu Feng、Simeng Sun、Runsen Feng、Tianyu He 和 Zhibo Chen。语义视频编码：将静态—动态线索注入面向 AI 任务的结构化比特流。arXiv 预印本 arXiv:2201.10162，2022。3[36] Nick Johnston、Damien Vincent、David Minnen、Michele Covell、Saurabh Singh、Troy Chinen、Sung Jin Hwang、Joel Shor 和 George Toderici。通过预激活和空间自适应比特率改进循环网络的有损图像压缩。载于 CVPR，第 4385–4393 页，2018。1[37] Aggelos K Katsaggelos、Lisimachos P Kondi、Fabian W Meier、J¨orn Ostermann 和 Guido M Schuster。MPEG-4 与基于率失真的形状编码技术。IEEE 会刊，86(6)：1126–1154，1998。3[38] Alexander Kirillov、Kaiming He、Ross Girshick、Carsten Rother 和 Piotr Doll´ar。全景分割。载于 CVPR，第 9404–9413 页，2019。8[39] Eastman Kodak。Kodak 无损真彩色图像套件（photoCD pcd0992）。http://r0k.us/graphics/kodak，1993。6[40] Nam Le、Honglei Zhang、Francesco Cricri、Ramin Ghaznavi-Youvalari 和 Esa Rahtu。面向机器的图像编码：一种端到端学习方法。载于 ICASSP 2021—2021 IEEE 声学、语音与信号处理国际会议（ICASSP），第 1590–1594 页。IEEE，2021。1, 3, 12[41] Nam Le、Honglei Zhang、Francesco Cricri、Ramin Ghaznavi-Youvalari、Hamed Rezazadegan Tavakoli 和 Esa Rahtu。面向机器的学习式图像编码：一种内容自适应方法。载于 2021 IEEE 多媒体与博览会国际会议（ICME），第 1–6 页。IEEE，2021。1[42] Mu Li、Wangmeng Zuo、Shuhang Gu、Jane You 和 David Zhang。学习内容加权的深度图像压缩。IEEE模式分析与机器智能汇刊，2020。1[43] Mu Li、Wangmeng Zuo、Shuhang Gu、Debin Zhao 和 David Zhang。学习用于内容加权图像压缩的卷积网络。载于 CVPR，第 3214–3223 页，2018。1 – .

[44] Xin Li、Jun Shi 和 Zhibo Chen。通过强化学习实现任务驱动的语义编码。IEEE图像处理汇刊，2021。3[45] Yanghao Li、Hanzi Mao、Ross Girshick 和 Kaiming He。探索用于目标检测的纯视觉Transformer骨干网络。ECCV，2022。1, 3[46] Tsung-Yi Lin、Piotr Doll´ar、Ross Girshick、Kaiming He、Bharath Hariharan 和 Serge Belongie。用于目标检测的特征金字塔网络。载于 CVPR，第 2117–2125 页，2017 年。1, 3。[47] Tsung-Yi Lin、Michael Maire、Serge Belongie、James Hays、Pietro Perona、Deva Ramanan、Piotr Doll´ar 和 C Lawrence Zitnick。Microsoft COCO：上下文中的常见对象。载于 ECCV，第 740–755 页。Springer，2014。5, 6, 12[48] Shu Liu、Lu Qi、Haifang Qin、Jianping Shi 和 Jiaya Jia。用于实例分割的路径聚合网络。载于 CVPR，第 8759–8768 页，2018。1, 3[49] Ze Liu、Yutong Lin、Yue Cao、Han Hu、Yixuan Wei、Zheng Zhang、Stephen Lin 和 Baining Guo。Swin Transformer：使用移位窗口的层次化视觉Transformer。载于 ICCV，第 10012–10022 页，2021。1, 2, 3, 4[50] Jonathan Long、Evan Shelhamer 和 Trevor Darrell。用于语义分割的全卷积网络。载于 CVPR，第 3431–3440 页，2015。1, 3, 8[51] Siwei Ma、Xiang Zhang、Shiqi Wang、Xinfeng Zhang、Chuanmin Jia 和 Shanshe Wang。联合特征与纹理编码：通过前端智能实现智能视频表示。TCSVT，29(10):3095–3105，2018。1, 3[52] Fabian Mentzer、Eirikur Agustsson、Michael Tschannen、Radu Timofte 和 Luc Van Gool。用于深度图像压缩的条件概率模型。载于 CVPR，第 4394–4402 页，2018。1[53] Fabian Mentzer、George Toderici、Michael Tschannen 和 Eirikur Agustsson。高保真生成式图像压缩。arXiv 预印本 arXiv:2006.09965，2020。1, 3[54] David Minnen、Johannes Ball´e 和 George Toderici。用于学习式图像压缩的联合自回归与层次先验。载于 NeurIPS，2018。1, 2, 5, 6[55] David Minnen 和 Saurabh Singh。用于学习式图像压缩的逐通道自回归熵模型。载于 2020年IEEE国际图像处理会议 (ICIP)，第 3339–3343 页。IEEE，2020。2, 5[56] Majid Rabbani 和 Rajan Joshi。JPEG 2000 静态图像压缩标准概述。《信号处理：图像通信》，17(1):3–48，2002。1, 2[57] Joseph Redmon、Santosh Divvala、Ross Girshick 和 Ali Farhadi。只看一次：统一的实时目标检测。载于 CVPR，第 779–788 页，2016。1, 3[58] Joseph Redmon 和 Ali Farhadi。YOLO9000：更好、更快、更强。载于 CVPR，第 7263–7271 页，2017。1, 3[59] Shaoqing Ren、Kaiming He、Ross Girshick 和 Jian Sun。Faster R-CNN：借助区域提议网络实现实时目标检测。NeurIPS，28:91–99，2015。1, 3[60] Oren Rippel 和 Lubomir Bourdev。实时自适应图像压缩。载于 ICML，第 2922–2930 页。PMLR，2017。3[61] Thomas Sikora。MPEG-4 视频标准验证模型。TCSVT，7(1):19–31，1997。3

[62] Saurabh Singh、Sami Abu-El-Haija、Nick Johnston、Jo-hannes Ball´e、Abhinav Shrivastava 和 George Toderici。端到端学习可压缩特征。载于 2020年IEEE国际图像处理会议 (ICIP)，第 3349–3353 页。IEEE，2020。1, 3[63] Gary J Sullivan、Jens-Rainer Ohm、Woo-Jin Han 和 Thomas Wiegand。高效视频编码 (HEVC) 标准概述。TCSVT，22(12):1649–1668，2012。1, 2, 6[64] Simeng Sun、Tianyu He 和 Zhibo Chen。面向多种智能应用的语义结构化图像编码框架。TCSVT，2020。2, 3, 6, 7[65] Lucas Theis、Wenzhe Shi、Andrew Cunningham 和 Ferenc Husz´ar。使用压缩自编码器进行有损图像压缩。arXiv 预印本 arXiv:1703.00395，2017。2[66] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko-reit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention is All You Need。NeurIPS，30，2017。3, 4[67] Gregory K Wallace。JPEG 静态图像压缩标准。IEEE消费电子学报，38(1):xviii–xxxiv，1992。1, 2[68] Shurun Wang、Shiqi Wang、Wenhan Yang、Xinfeng Zhang、Shanshe Wang、Siwei Ma 和 Wen Gao。通过可扩展特征与纹理压缩，构建便于分析的人脸表征。TMM，2021。3[69] Thomas Wiegand、Gary J Sullivan、Gisle Bjontegaard 和 Ajay Luthra。H.264/AVC 视频编码标准概述。TCSVT，13(7):560–576，2003。1[70] Yuxin Wu、Alexander Kirillov、Francisco Massa、Wan-Yen Lo 和 Ross Girshick。Detectron2。https://github.com/facebookresearch/detectron2，2019。6[71] Enze Xie、Wenhai Wang、Zhiding Yu、Anima Anandkumar、Jose M Alvarez 和 Ping Luo。SegFormer：采用Transformer的简洁高效语义分割设计。NeurIPS，34:12077–12090，2021。1, 3, 8[72] Yibo Yang、Robert Bamler 和 Stephan Mandt。改进神经图像压缩的推理。NeurIPS，33:573–584，2020。2[73] Hengshuang Zhao、Jianping Shi、Xiaojuan Qi、Xiaogang Wang 和 Jiaya Jia。金字塔场景解析网络。载于 CVPR，第 2881–2890 页，2017。8[74] Rui Zhao、Wanli Ouyang、Hongsheng Li 和 Xiaogang Wang。通过多上下文深度学习进行显著性检测。载于 CVPR，第 1265–1274 页，2015。8[75] Sixiao Zheng、Jiachen Lu、Hengshuang Zhao、Xiatian Zhu、Zekun Luo、Yabiao Wang、Yanwei Fu、Jianfeng Feng、Tao Xiang、Philip HS Torr 等。借助Transformer从序列到序列视角重新思考语义分割。载于 CVPR，第 6881–6890 页，2021。1, 3, 8[76] Lei Zhou、Zhenhong Sun、Xiangji Wu 和 Junmin Wu。采用注意力机制的端到端优化图像压缩。载于 CVPR研讨会，第 0 页，2019。5[77] Yinhao Zhu、Yang Yang 和 Taco Cohen。基于Transformer的变换编码。载于 ICLR，2022。1, 2, 4, 6

本补充材料提供了更多实验对比、语义感知加密的实现细节，以及将所有前景对象作为感兴趣区域的消融研究。

A. 与其他方法的比较

本节将我们提出的方法与 VTM 18.2 以及另外三种面向机器的图像编码方法进行比较：基于传统编解码器的 RoI 比特分配方案 [33]，基于学习的联合训练编解码器 [40]，以及基于通用表征学习的方法 [21]。评估任务是在 COCO 上进行实例分割 2017 [47]。值得注意的是，RoI 比特分配方法的比特流 [33]和任务驱动的联合训练方法均针对相应任务进行了定制，而基于通用表征学习的方法 [21] 需要使用新的源数据（即学习到的表征）重新训练任务模型。尽管如此，与其他方法相比，我们的方法仍然取得了显著的改进。

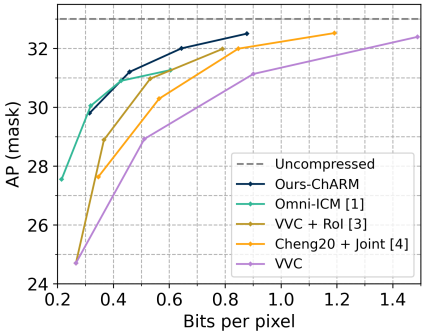
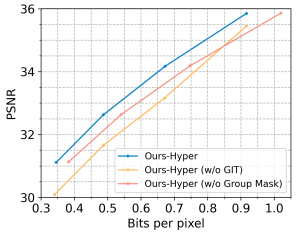


图 13. MS COCO 上的实例分割。

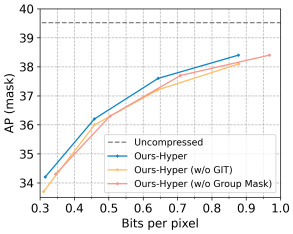
B. 语义感知加密

我们对潜变量进行加密。具体而言，对于每个待加密的组，首先将选定的变量重塑为一维向量，然后使用 Fisher-Yates 洗牌算法 [22]对其进行扰乱。扰乱通过加密随机种子实现。接着，将扰乱后的向量恢复为原始形状，并编码为相应的比特流。相应地，似然也以相同的方式打乱，以进行熵编码。随后，对当前组的潜变量进行熵编码，生成加密比特流。最后，根据需求，语义结构化比特流由加密和未加密的比特流组成。此外，用户可以对不同组进行不同级别的加密，从而生成分层加密比特流。

C. 消融研究



(a) 感兴趣区域重建。



(b) 实例分割。

图14. 消融研究。

在正文第 5 节中，我们将人视为感兴趣区域（RoI），进行了消融研究。本节中，我们提供了补充实验结果，并将属于 MS COCO 类别的所有前景对象视为 RoI。图14 所示结果表明，这与正文第 5 节的结论相同：我们提出的组掩码可以显著节省比特率，而组独立变换能够确保选择性传输和重建的重建质量。