机器感知驱动的图像压缩：分层生成方法

Yuefeng Zhang、Chuanmin Jia、Jianhui Chang、Siwei Ma ，*IEEE* 高级会员

arXiv:2304.06896v1 [eess.IV] 14 Apr 2023

摘要——**在这个信息时代，图像是存储和传输信息的重要媒介。随着图像数据量迅速增长，视觉压缩和视觉数据感知成为备受关注的两个重要研究课题。然而，这两个课题鲜少被一同探讨，且各自沿着不同的研究路径发展。基于学习的图像压缩方法提供了紧凑的压缩域表示，因此有望通过单一数据流，同时实现高效的数据存储与压缩，以及机器感知任务。本文提出一种分层生成式图像压缩模型，即使在极高压缩比下，也能获得面向人类视觉的高质量重建图像。为提高分析效率和灵活性，本文提出一种任务无关的基于学习的压缩模型。与传统编解码器和基于学习的编解码器相比，该模型在有效支持各种基于压缩域的分析任务的同时，仍能保有出色的重建感知质量。此外，本文采用联合优化方案，以在压缩比、重建图像质量和下游感知性能之间取得最佳平衡。实验结果表明，本文提出的基于压缩域的多任务分析方法能够取得与基于RGB图像的方法相当的分析结果，同时节省最高99.6%的比特率（即与将原始RGB图像作为分析模型输入相比）。从模型大小和信息保真度两方面，进一步验证了该模型的实用能力。**

索引词—**图像压缩、机器感知、生成对抗网络、视觉应用。**

I. 引言

VISUAL 由于电子图像采集领域取得了巨大进步，视觉数据正以惊人的速度生成。一方面，我们需要数据压缩，以便高效存储和传输海量视觉数据。另一方面，要将像素级信号转化为面向机器的知识，则需要高层语义理解。过去，由于这两个领域的目标不同，它们很少被同时研究 [1]。

视觉数据量不断增长，这凸显了对存储和传输以及机器感知的需求。除了为人类视觉有效重建像素这一目标外，图像/视频编码标准也开始关注高层机器感知和模式识别任务。为支持大规模图像/视频检索与分析，MPEG推出了标准：用于视觉搜索的紧凑描述符（CDVS）（ISO/IEC15938-13） [2]，于2015年9月发布；以及视频分析紧凑描述符（CDVA）（ISO/IEC15938-14） [3] ，于2019年7月发布。与此同时，JPEG AI [4]于2019年3月启动，旨在满足人类视觉和计算机视觉任务的需求。

面向机器的视频编码（VCM） [5] 或智能编码[6]旨在构建高效的联合视觉压缩与分析框架。这些方法可分为两类。第一类 [7], [8] 旨在使重建图像更适用于面向机器的任务。该类中的大多数方法都是基于传统编解码器构建的，通过引入分析相关信息来实现这一目标。第二类 [9], [10] 直接对中间压缩特征进行视觉分析。在这一类别中，基于学习的压缩方法具有优势，因为其深度特征包含丰富的紧凑语义信息，而传统编解码器并不具备这一特性。

得益于深度学习，基于学习的方法在率失真效率方面开始超越传统压缩技术 [11], [12]。与此同时，视觉应用也正在经历巨大变革，例如图像检测 [13],、识别 [14],和分割 [15]任务。因此，这为将压缩与视觉分析相结合提供了可能性。在图像压缩中使用基于学习的方法的一个优势在于，其压缩域是学习得到的，因此包含丰富的紧凑语义信息，而这是传统编解码器所不具备的。直观而言，压缩数据旨在重建原始图像，因此可以将其理解为一种高度紧凑的视觉表征[9], ，这意味着直接在压缩域上执行高级图像处理和机器感知任务是可行的。

疫情之后，各组织开始鼓励员工通过视频会议与同事互动，以实现在家办公；而在居家隔离期间，短视频平台则帮助人们消磨时间。所有这些因素共同推动了视觉数据的爆炸式增长。如何以高效且经济的方式存储和传输这些数据，已引起业界的广泛关注。Facebook 宣布了其视频聊天技术方案 [16] ，通过传输人脸关键点并在低带宽条件下使用生成模型来实现视频聊天。高通发布了实时 1280×704 视频解码演示 [17] ，展示了搭载高通骁龙888 处理器的智能手机上的解码效果。在他们的工作中，神经视频解码由人工智能推理引擎和可并行化的熵编码提供支持。鉴于当前对带宽的巨大需求，视觉完整性在极高压缩比条件下于实际应用中的重要性愈发凸显。

为弥合以往研究与实际应用之间的差距，我们聚焦于人脸数据，并提出一种分层生成式压缩模型，即使在极高压缩比下也能保持视觉保真度。

分层端到端压缩流水线的设计

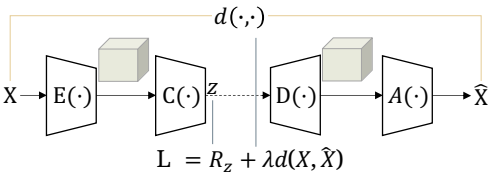

图 1. 图像压缩框架。

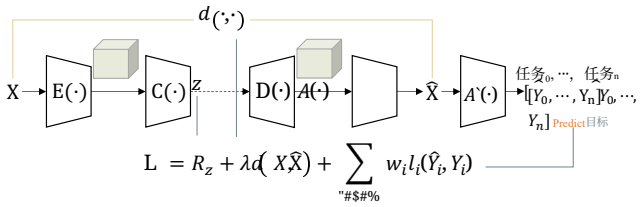

图 2. 在重建图像上进行机器感知的框架。

其根本动机源于 Marr 的计算理论[18] ，该理论认为，视觉可以被理解为一个信号处理系统，而这一观点天然蕴含了对分层操作的支持。根据图像到图像系列研究 [19]–[22]，图像风格和纹理信息可以用一维向量描述，而内容信息仍然需要空间信息，其表示形式为二维。此外，研究证明，通过分层操作可以有效保留关键信息，尤其是在极低比特率编码场景下 [23], [24]。

与此同时，为了直接分析用于机器感知任务的压缩数据，提出了一种任务无关的多任务分析模型。

本文的主要贡献可概括如下：

1）我们研究并从理论上分析了机器感知驱动的基于学习的压缩模型框架的特性。通过这些分析，我们提出了这些方法的优化公式，并揭示了现有分层图像压缩方法缺乏灵活性的问题。

2）我们提出了一种机器感知驱动的分层压缩模型，该模型采用端到端的方式进行学习，使每一层都具有可学习性，因而富含语义信息。所提出的模型同时适用于人类视觉和机器感知任务，并在 CelebAMask-HQ 数据集上进行了大量实验。

3）就编码性能而言，所提出的模型在感知指标（即 FID [25]，DISTS [26],和 LPIPS [27]）上优于传统编解码器（例如 VVC）和最先进的基于学习的编解码器，尤其是在极低比特率条件下（即每像素比特数≤ 0.1）。对于语义分析任务，在压缩数据上实现了可比的视觉分析性能；与分析原始 RGB 图像相比，传输比特率最多可节省 99.6%。此外，提出的多任务分析模型可进一步提升视觉任务的性能并降低模型复杂度。

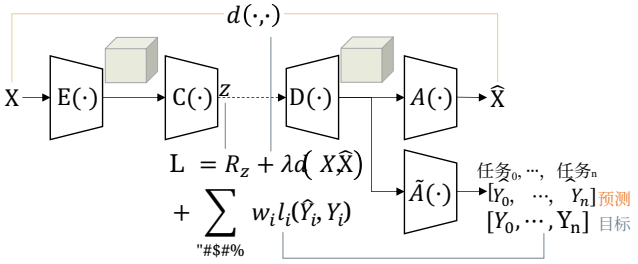

图 3. 直接对压缩视觉数据进行机器感知的框架。

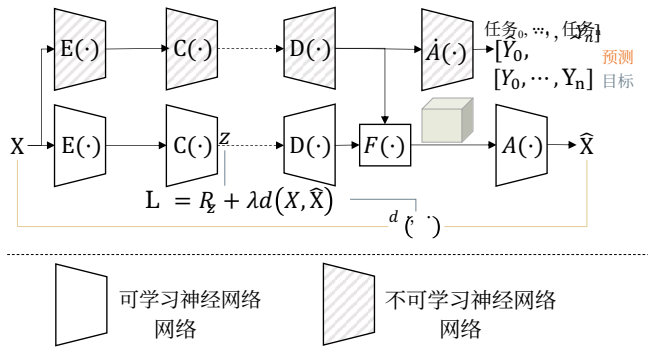

图4. 边信息辅助的分层图像压缩框架。

本文其余部分安排如下。第二节分析机器感知编码及其优化形式。第三节回顾本文的相关工作。第四节通过展示框架介绍所提出方法的主要贡献。第五节和第六节基于我们的模型展示实验结果并进行讨论。最后，第七节讨论结论及对未来工作的启示。

II. 问题表述

为说明机器感知驱动的压缩模型的一般设置，我们将以下组件定义为 [1]:

- E($\cdot|\theta_E$)，将视觉输入变换为特征矩阵；
- C($\cdot|\theta_C$)，将特征编码为比特流，包括概率估计和熵编码步骤；
- D($\cdot|\theta_D$)，将比特流解码为特征矩阵。

针对不同的输入数据类型和用途，我们定义以下分析网络：

- A($\cdot|\theta_A$)，用于将特征投影为供人类视觉使用的重建图像；
- A′($\cdot|\theta_{A'}$)机器视觉任务服务器，以重建图像作为输入；.
- 面向机器视觉任务的 A($\cdot|\theta_{\cdot A}$) 服务器，以手工设计的侧特征作为输入；~
- A($\cdot|\theta_{\sim A}$) 机器视觉任务服务器，将中间解码特征作为输入。

我们根据训练目标对以往的相关研究进行分类，并分析其对应的训练组件。

在梯度反向传播中。对于图 1 所示的基于学习的图像压缩方法，其训练目标是平衡码率与失真之间的权衡：

$$\arg\min_{\Theta} \mathrm{R}_z + \lambda \mathrm{d}, \quad \Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{A}}\}. \tag{1}$$

考虑到机器感知，训练目标是找到码率、失真和视觉任务性能之间的平衡点，由此可将优化函数表示为：

$$\arg\min_{\Theta} \mathrm{R}_z + \lambda d(X, \hat{\mathrm{X}}) + \sum_{0 \le i \le n} w_i l_i(\hat{Y}_i, Y_i), \tag{2}$$

ˆ其中， X 是输入图像， X 是重建图像， Y_i 是第 i 个下游任务的真实标签， Y_i 是任务 i的预测结果， d 是图像失真度量， l_i 表示任务 i的任务损失度量。如果下游任务以重建图像 X 作为输入，如图2所示，则视觉任务的预测结果 Y 和优化参数为：

$$\hat{Y} = \mathrm{A}'(\hat{X}), \quad \Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{A}}, \theta_{\mathrm{A}'}\}. \tag{3}$$

如果分析模型以中间特征 Z 作为输入，如图 3 所示，则优化函数为：

$$\hat{Y} = \widetilde{\mathrm{A}}(Z), \quad \Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{A}}, \theta_{\widetilde{\mathrm{A}}}\}. \tag{4}$$

对于分离训练，分析模块通过优化 $\theta_{\mathrm{A}'}$ 或 $\theta_{\widetilde{\mathrm{A}}}$进行训练，且独立于压缩模块。

III. 相关工作

A. 用于机器感知的压缩视觉数据

关于如何分析压缩表示的已有研究分为两类：传统编解码器压缩表示和基于学习的编解码器压缩表示。

从 JPEG [28] 到最新的 VVC [29] 标准，传统编解码器均可视为基于块的预测、变换和量化。为了直接分析传统编解码器生成的图像/视频比特流，通常会将变换参数 [7], [8], [30] 或中间编码块 [31], [32] 作为输入。然而，与学习方法相比，这些压缩表示均由人工设计，缺乏可扩展性。

基于学习的编解码器 [11], [33], [34] 通常采用自动编码器架构，直接优化率失真权衡，并通过量化、概率估计和熵编码实现高压缩比。基于变分自编码器（VAE）的压缩基本框架如图 1 所示，其中使用像素级失真损失进行训练。随着机器视觉日益受到关注，研究人员开展了相关工作 [35]–[37] ，以探究当分析模型的输入是重建图像而非原图时，压缩如何影响机器感知性能。图 2 展示了这些方法的高层视图：重建图像 X 被输入视觉分析模型 A'（·）。

为了高效执行机器感知任务，鉴于基于学习的压缩方法所产生的压缩表示具有灵活且富含语义的特点，近期研究 [1], [9], [10], [38]已证明，无需解码即可直接在压缩表示上执行视觉任务具有潜力，其框架如图 3 所示。根据对任务相关先验知识的需求，这些方法可分为两类：任务感知和任务无关。对于第一类，压缩模型通过联合训练压缩模型和任务分析模型来解决特定的下游任务 [1], [6], [10], [39] 。对于第二类，压缩过程不考虑待解决的下游任务 [9], [40]。

B. 分层图像压缩

分层图像压缩得益于这样一种先验知识：预定义信息可以采用多种形式，例如边缘图 [23], 语义分割 [41],人脸关键点[16],等。如图4所示，视觉数据首先被变换为分层特征分量，其中一个由人工设计，另一个通过学习获得，然后分别进行编码。预定义的层信息旨在辅助特定的下游视觉分析任务，例如，人脸边缘图可以支持关键点检测 [42]。

IV. 方法

我们提出的模型的整体流程图如图 5 所示，包括压缩部分和机器任务分析部分。

A. 基于学习的分层压缩

我们的基于学习的分层压缩模型包含三个主要组成部分：编码器、解码器和概率估计模型。我们首先设计了一个分层编码器和一个生成式解码器，以高效压缩视觉数据。随后概述了优化目标。

1) 分层编码器：根据图像到图像系列研究 [19]–[22]，，图像风格和纹理信息可以用一维向量描述，而内容信息仍需要空间信息，其表示形式为二维。此外，研究证明，通过分层操作可以有效保留关键信息，尤其是在极低比特率编码场景下 [23]–[24]。

采用不同的编码器将输入图像信号 X 编码为特征 I_R 和 I_S，分别用于重建和语义分析。将相应的编码器记为 E_R 和 E_S ，则可计算出潜在表示：

$$I_R = E_R(X), I_S = E_S(X). \tag{5}$$

对于不同的数据分布，我们使用独立的概率分布模型进一步将 I 变换为紧凑的 z，随后采用基于数据分布估计的熵编码方法将其编码为比特流。通过计算交叉熵 $H(p, q)$，这一点已得到证明。

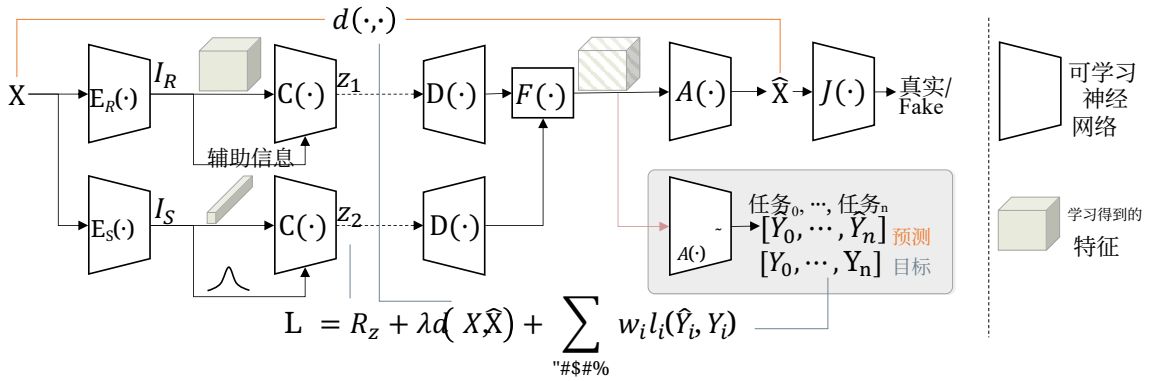

图 5. 所提方法概览。我们将原始图像信号变换为双层表示：面向重建的 I_R 和面向语义的 I_S。图中展示了两个目标，包括：1）压缩；2）对压缩表示进行机器感知。

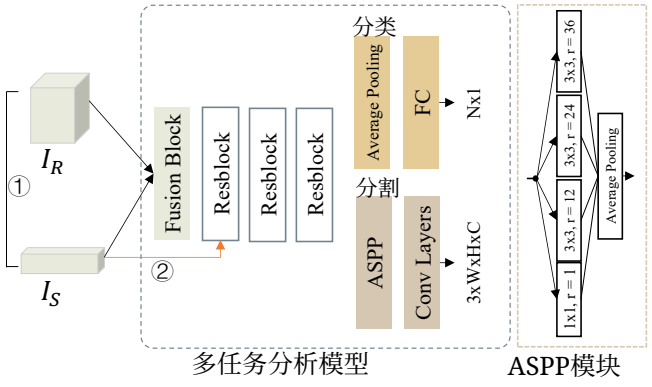

图 6. 多任务分析网络架构示意图。这里， C 和 S 分别表示分类和分割的类别数。对于残差块，我们采用了 [43]中第二种深度残差块设计。

作为 z理论比特流大小的上界， [33],采用参数化概率模型 q ，以间接估计 z 的概率分布，其概率分布为 p。

具体而言，超先验熵模型 [33] 用于对 I_R进行熵估计。潜在表示 I_R首先被量化为由离散符号组成的向量$\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$。将超先验记为$\mathbf{Y}$，条件概率可写为 $P_{\mathbf{X}|\mathbf{Y}}$。$\mathbf{X}$ 中的每个元素 X_i 的条件分布可假设为高斯分布，即:

$$p_{\hat{\boldsymbol{X}}|\hat{\boldsymbol{Y}}}(\hat{\boldsymbol{X}} \mid \hat{\boldsymbol{Y}}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{6}$$

潜在表示的似然可按如下方式计算:

$$\begin{aligned} p_{\hat{\boldsymbol{X}}|\hat{\boldsymbol{Y}}}(\hat{\boldsymbol{X}} = \hat{x}_i \mid \hat{\boldsymbol{Y}}) &= \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) * \mathcal{U}(-\frac{1}{2}, \frac{1}{2})(\hat{x}_i) \\ &= \phi(\hat{x}_i + \frac{1}{2}) - \phi(\hat{y}_i - \frac{1}{2}) \end{aligned} \tag{7}$$

其中， ϕ 表示标准高斯分布的累积分布函数。鉴于其数据维度较低， I_S 采用了具有高斯分布的非参数化因子分解密度模型[11] 。

2）融合模块和解码器：我们的解码器在融合模块 F 的帮助下，以生成式的方式重建原始视觉信号。采用自适应实例归一化（AdaIN） [19] 作为融合模块，其中，富含语义信息的表征 z_2 可视为多层感知机（MLP）的输入，用于预测残差块中卷积层的均值和方差，即 $F(z_1, z_2) = \text{AdaIN}(z_1, \gamma, \beta) = \gamma\left(\frac{z_1 - \mu(z_1)}{\sigma(z_1)}\right) + \beta$，其中， μ， σ 分别表示通道均值和标准差，而 γ， β则由以 z_2 为输入的多层感知机学习得到。解码器采用对抗损失，以生成符合人眼视觉感知的重建图像。

3）优化目标：训练设置对于连接数据压缩与视觉分析至关重要。根据任务目标，训练策略分为两类：a）**面向人类视觉任务**；b）**面向机器感知任务**。我们首先介绍人类视觉任务的训练设置。然后，在下一小节中，我们将讨论机器感知任务的设置。

对于人类视觉任务，我们模型的目标是通过预测像素值来重建原始图像信号，同时最小化 X 与 X 之间的失真。为了在保留信号保真度的同时满足人类视觉需求，我们的失真度量包含三个部分：逐像素平均绝对误差（MAE）损失 d_{MAE}、基于整体结构的SSIM损失（即 $d_{SSIM} = 1 - SSIM(X, X)$），以及感知损失 d_p ，如 [44]所示。失真度量 d 可表示如下:

$$d(X, \hat{X}) = \lambda_{MAE} d_{MAE} + \lambda_{SSIM} d_{SSIM} + \lambda_p d_p, \tag{8}$$

其中， λ_{MAE}、 λ_{SSIM}和 λ_p 为超参数。对于编码器和解码器部分，训练优化方式如下:

$$\mathcal{L}_1 = \mathbb{E}_X[-\lambda \log(R(\hat{z})) + d(X, \hat{X}) - \beta \log(\mathrm{J}(\hat{X}, z_1))], \tag{9}$$

$$\Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{F}}, \theta_{\mathrm{A}}\}, \tag{10}$$

其中，J(·) 是判别器，它通过在训练过程中判断重建图像是真实图像还是伪造图像，帮助生成/解码满足感知要求的图像。J(·) 的训练损失可定义为:

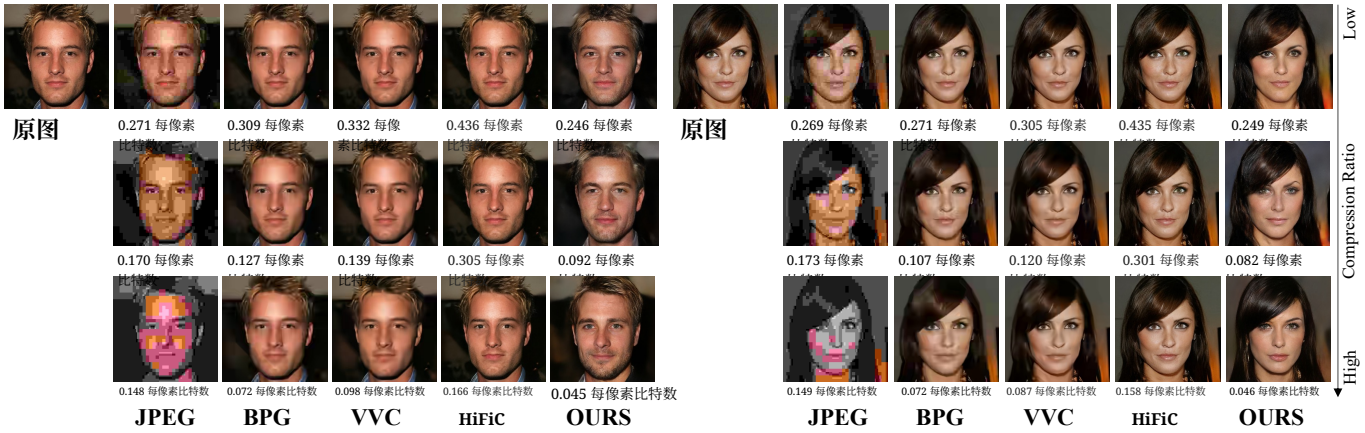

图7. 传统编解码器和基于学习的编解码器重建图像的视觉比较。对于每幅重建图像，我们在图像下方显示其比特率（即比特每像素，bpp）。由于压缩模型得到的比特率并非连续值，我们采用二分搜索，为每种压缩方法找到与我们的方法输出比特率最接近的结果。

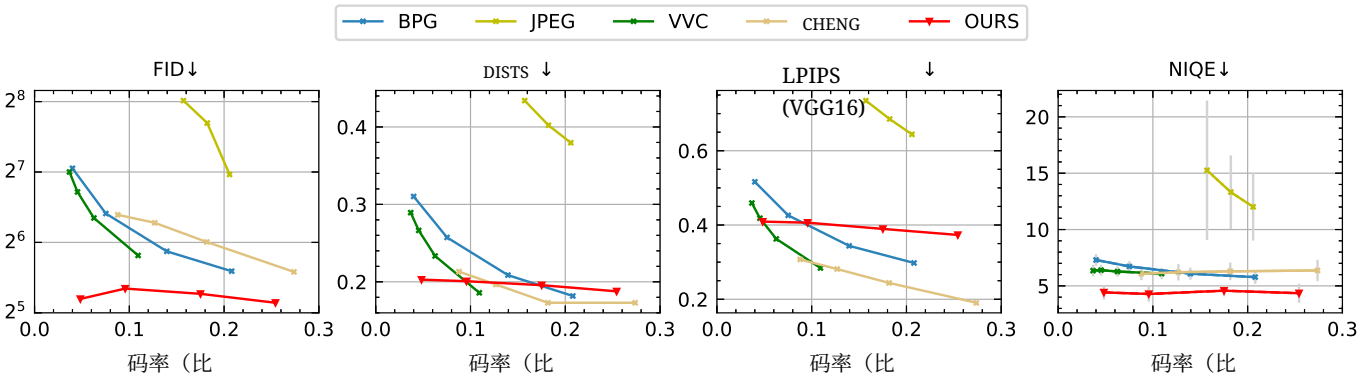

图8. 我们的压缩模型的率感知曲线。标题中的箭头 ↓ 表示该指标值越低越好。NIQE指标中的灰色箭头条表示数据方差，而其他三个指标的数据方差很小，可以忽略不计。

$$\mathcal{L}_{\mathrm{J}} = \mathbb{E}_X[-\log(1-\mathrm{J}(\hat{X}, z_1))] + \mathbb{E}_X[-\log(\mathrm{J}(X, z_1))], \tag{11}$$

其中， $\mathcal{L}_{\mathrm{J}}$ 是判别器的损失函数。需要注意的是，判别器 J 以 z_1 为条件，例如 [44] ，以获得清晰的图像；我们采用 [45] 提出的多尺度判别器方法来消除模式崩溃问题。判别器 J 的网络设计细节见表 II。

在此目标设置下，训练过程中会交替优化参数 $\Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{F}}, \theta_{\mathrm{A}}\}$ 和 $\Theta = \{\theta_{\mathrm{J}}\}$。为了在量化过程中使用梯度下降方法，训练时会添加均匀噪声；在推理阶段则采用舍入操作，将数值舍入到最接近的整数，遵循 [11]。机器视觉任务的优化目标设置将在第四节 B 部分讨论。

B. 压缩数据上的机器感知

基于学习的图像压缩模型凭借灵活且经过学习的潜在表示，有望弥合视觉数据压缩与机器感知之间的差距。接下来将介绍以压缩数据为输入的视觉分析，以及优化目标的设置。

1) 多任务分析： 我们的假设是，核心语义已嵌入并学习于 $\hat{}z_1$ 和 $\hat{}z_2$ 中，因此无需解码即可解决多个下游任务。为了进一步提高分析效率，我们从多任务分析的角度处理多个任务，因为，

$$\hat{Y} = \widetilde{\mathrm{A}}(z_1, z_2), \quad \hat{Y} = [\hat{Y}_0, \cdots, \hat{Y}_n], \tag{12}$$

$\hat{}$ 其中， n 是下游任务的总数， Y_i 是第 i个任务的预测。分类和分割任务被列为示例，因为它们能够代表机器感知任务，且已得到广泛研究，不会损失通用性。多任务分析网络的详细信息如图 6所示，其中包含任务相关模块。多任务训练最重要的方面之一是学习如何平衡不同的任务损失 [46]。在此，我们设定超参数来控制各任务之间的权衡，因此多任务视觉分析损失可表示为：

$$\mathcal{L}_{\widetilde{\mathrm{A}}} = \lambda_{cls} l_{cls} + \lambda_{seg} l_{seg}, \tag{13}$$

其中，l_{cls} 和 l_{seg} 分别是与分类和分割相关的任务损失，$\lambda_{cls}, \lambda_{seg}$ 分别是它们对应的权重超参数。在消融研究中，我们通过实验证讨论了这些超参数的选择。

2）优化目标：我们研究了两种训练流程，以进一步探索压缩与机器感知之间的关系，从而针对机器感知任务优化压缩模型：分离训练和联合训练。我们考虑了需要优化的不同参数。

在分离训练设置中，我们固定压缩模型，仅训练多任务分析模型，即 $\Theta = \{\theta_{\widetilde{\mathrm{A}}}\}$。对于联合训练，压缩模型和多任务分析模型通过总损失及相应参数进行联合优化：

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{\mathrm{J}} + \gamma \mathcal{L}_{\widetilde{\mathrm{A}}}, \tag{14}$$

$$\Theta = \{\theta_{\mathrm{E}}, \theta_{\mathrm{C}}, \theta_{\mathrm{D}}, \theta_{\mathrm{F}}, \theta_{\mathrm{A}}, \theta_{\widetilde{\mathrm{A}}}\}, \tag{15}$$

其中，超参数 γ 用于平衡人类视觉与机器感知目标之间的优化点。具体而言，参数 θ_{E}，θ_{C}，θ_{D}，θ_{F}，θ_{A} 是分离训练得到的微调结果，而非从头开始训练，后续遵循 [9]，并且我们在第五节中将γ 设为 1。

V. 实验

我们通过实验证明，基于学习的图像压缩方法的主要优势并非在于信号级保真，而在于实现高感知质量的重建和语义级信息保留，从而能够在压缩视觉数据上执行机器感知任务。

A. 实验设置

1）数据集：为评估所提出的模型，我们在 CelebAMask-HQ [47] 数据集上开展了大量实验，以研究该模型在人类视觉和机器感知方面的表现。CelebAMask-HQ 是一个大规模、高分辨率的面部语义数据集，包含 30,000 张分辨率为 512×512的人脸图像。每张图像均标注了 40 个属性类别，并具有 19 个语义类别的高质量像素级标签，即 18 个前景类别和 1 个背景类别。我们遵循 [48] 中的官方数据集划分设置，并报告模型在 256×256 图像上的表现。

2）图像质量指标：我们使用面向人类感知的图像质量评估指标（即 FID [25], DISTS [26], LPIPS [27]）评估重建图像。这些指标旨在根据模型在低层视觉任务中的感知表现对模型进行排名。我们使用在 ImageNet [49]上预训练的 VGG16 c 网络作为 LPIPS 评估的特征提取器。此外，我们还采用无参考指标 NIQE [50]，该指标通过统计量衡量重建图像分布偏离自然图像分布的程度。

3) 机器感知指标：我们通过两项任务验证直接分析压缩域的可行性以及多任务训练的有效性：a) **多属性估计**通过计算 40 个属性的平均预测准确率（准确率）进行评估；b) **语义分割** 通过所有区域的像素准确率、每种语义类别的类别准确率进行评估。类别、通过对全部 19 个类别取平均计算的平均交并比（mIoU），以及在 mIoU 的基础上纳入类别出现频率考量的频率加权交并比（FWIoU）。

4）对比编解码器：我们将本方法重建图像的质量与以下传统和基于学习的编解码器基线进行比较。我们选取三种传统编解码器（即 JPEG、BPG、VVC）作为对比方法。对于 JPEG，我们使用 cjpeg[1] 实现。BPG 是基于 H.265/HEVC 标准的开源软件[2] 。VTM[3] 11.0 版是 H.266/VVC 的参考软件。对于 BPG 和 VVC，我们均使用 YUV420 色彩空间对图像进行编码。对于基于学习的压缩方法，我们选取基于生成式解码的 HiFiC [44]，以及在主观指标峰值信噪比（PSNR）方面达到与 VVC 相当性能的 CHENG [12] 进行比较。

B. 实现细节

1）独立训练计划：压缩部分和机器感知部分分别采用独立的训练计划。对于压缩部分，分别优化优化目标式（10）和 $\Theta = \{\theta_{\widetilde{\mathrm{A}}}\}$。压缩部分使用 Adam [51] 作为优化器，其中 β_1 和 β_2 采用默认值 0.9 和 0.999。我们采用批次大小为 8，并将学习率固定为 0.0002。我们遵循最小二乘生成对抗网络（LSGAN）训练策略 [52]，并采用双时间尺度更新规则（TTUR）[25]，为判别器和解码器/生成器设置独立的学习率，以帮助训练收敛。我们从头开始训练压缩模型，共进行 3.6M 次迭代。式（8）中的 λ_{MAE}、λ_{SSIM} 和 λ_p 分别设为 10、0.25 和 0.2。

对于机器感知部分，在单独训练中，我们对单任务和多任务分析网络采用相同的训练策略。模型使用随机梯度优化器进行优化，训练 50 轮次，小批次大小为 32。学习率从 0.001 开始，每 10 轮次乘以 0.1。值得注意的是，我们不采用图像增强方法（例如随机裁剪和翻转），因为这些方法在压缩域中没有语义意义。

2）联合训练计划：在联合训练阶段，我们通过式（14）中的目标损失联合优化压缩模型和多任务分析模型，其待优化参数如式（15）所示。我们使用单独训练得到的预训练检查点对式（15）进行微调，训练 10 轮次，学习率固定为 0.0001。

3）码率控制：鉴于我们的设计中采用了融合模块 F(·)，我们提出了一种简单直接的码率控制策略，以获得不同的压缩比。观察图 10(a) 后，我们发现，近 80% 的潜变量通道对估计码率的贡献很小。此外，如热力图图 10(b) 所示，通道之间的关联性相当弱，因此我们可以仅利用部分特征通道来表示潜在表示。基于这一分析，我们开展了实验。

[1]https://linux.die.net/man/1/cjpeg [2] https://bellard.org/bpg/ [3] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware VTM

表 I 多任务分析网络的分类和分割结果。* 表示我们考虑每通道8 位的 RGB 图像，总计每像素24 位（每像素比特数）。

输入类型	编解码器	Rate（每像素比特数）	分割				分类	质量
			mIoU	FwIoU	准确率	分类准确率	准确率	DISTS ↓
原图 RGB	/	24*	0.692	0.876	93.29%	78.04%	90.59%	-
重建图 RGB	Low	0.096	0.591	0.814	89.58%	69.79%	88.12%	0.201
已压缩 Data	提取。	0.048	0.557	0.790	88.07%	65.90%	87.76%	0.203
	Low	0.096	0.572	0.805	89.01%	67.31%	88.01%	0.201
	中间	0.175	0.580	0.814	89.57%	69.73%	88.41%	0.196
	High	0.254	0.598	0.822	90.07%	69.73%	88.75%	0.188

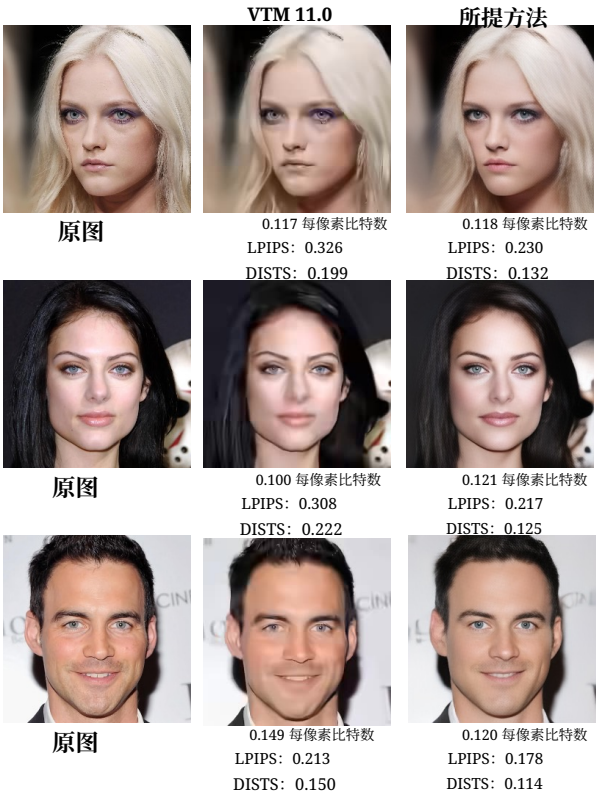

图 9. VVC（VTM 11.0，qp $= 37$）与我们的方法的定性比较结果。LPIPS 和 DISTS 值越低，表示质量越好。

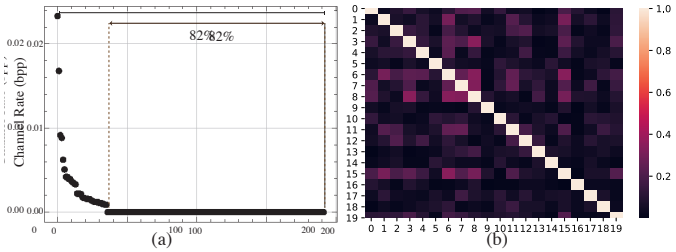

图 10. 逐通道可视化分析结果。(a) 按照每个特征通道所使用的码率进行降序排列。(b) 20 个随机选取的通道之间的皮尔逊相关系数的绝对值，其取值范围为 $[0, 1]$, ，数值越小表示相关性越弱。

表 II
压缩模型中判别器的架构示意图 $J(\cdot)$ 卷积层以“CONV”为前缀，后接通道数、卷积核尺寸、归一化和激活函数。↑ 和 ↓

IS 表示实例归一化，*LN* 表示层归一化。*LReLU*表示 Leaky ReLU 激活函数。
分别表示上采样和下采样。

判别器	
判别器1	判别器2
Conv64-4x4, ↓2, LReLU	Conv32-4x4, ↓2, LReLU
Conv128-4x4, ↓2, IS, LReLU	Conv64-4x4, ↓2, IS, LReLU
Conv256-4x4, ↓2, IS, LReLU	Conv128-4x4, ↓2, IS, LReLU
Conv512-4x4, IS, LReLU	Conv256-4x4、IS、LReLU
Conv1-4x4	Conv1-4x4

通过直接减少潜在表示的通道数量（即 Z）来获取具有实现极高压缩比潜力的压缩模型的码率控制策略。对于码率从极低到较高的压缩模型，我们依次将潜在表示的通道数量设置为集合 $\{8, 16, 64, 128\}$ 中的值。

C. 压缩性能

从主观评价的角度来看，我们在图 7 中展示了重建图像，以说明我们面向人类视觉的压缩模型即使在比特率限制极低的情况下（即图 7 最底行的图像），仍能保留有利于视觉感知的信息。显而易见，传统编解码器，包括 JPEG、BPG 和 VVC，表现不佳

存在严重的模糊和块效应。与 VVC 基线相比，所提出的方法以其三分之二或更低的比特率，生成了重建质量明显更高、纹理更细腻且更准确的图像。尽管 HiFiC 模型同样利用图像生成方法生成了视觉质量出色的图像，但它并未研究高压缩比的情况。

从客观评估的角度来看，旨在反映人类视觉感知的图像质量评估指标，包括 FID、DISTS、LPIPS 和 NIQE，如图 8 所示。尤其是在 FID 和 NIQE 指标上，我们的压缩模型大幅优于对比方法；在比特率极低的情况下（即每像素比特数 ≤ 0.1），我们的模型在 DISTS 和 LPIPS 指标上也展现出优势。客观而言，基于图中的结果

表 III 不同压缩比的压缩模型上多任务训练对分类带来的增益示意。 Δ 是多任务学习带来的增益。

	分类准确率（%）			
压缩模型	极端	Low	中间部分	High
单任务	86.74	87.74	87.76	87.82
多任务	87.76	88.01	88.41	88.75
Δ	+1.02	+0.27	+0.65	+0.93

（图 8），以 DISTS 指标为例，并以 VVC 为锚点，所提出的方法在等比特率条件下（即每像素比特数分别为 {0.2、0.4、0.6、0.8}）实现了平均重建质量提升，幅度分别为 39.1%、27.7%、15.1% 和 6.6%。

D. 机器感知性能

1）单任务与多任务： 在表 III 中，我们通过图 6 所示的多任务学习分析网络，探索了多任务学习带来的分类性能提升潜力。以分类任务为例，与单任务学习相比，引入多任务学习后准确率提升了 0。7%。

2）语义分析： 我们的多任务分析模型的语义分析结果见表 I。

这里我们采用四种指标评估分割任务。结果表明，不同指标反映了不同的退化程度。我们分析认为，这是因为我们的基于生成学习的压缩模型严重依赖先验知识（即数据分布），因此在出现罕见配件时表现较差，从而严重影响类别准确率等类别相关指标。

3）与*RGB*图像上的感知结果进行比较： 我们采用以 RGB 图像为输入的机器感知分析作为对比方法，并与我们的压缩潜变量输入模型进行比较。以我们的低比特率压缩模型为例，在表 I 中，与以原始RGB图像为输入相比，以其压缩潜变量为输入进行多任务分析时，分割任务的 mIoU 指标降低了 0。8%，同时节省了 99。6% 的比特率（即 $(24-0.096)/24\times100\%$）。观察表 I 所示的实验结果后，可以得出以下结论：

- 机器感知可以在压缩域上进行，且性能与输入 RGB 图像相当。
- 压缩比是以牺牲重建图像质量和机器感知性能为代价获得的。
- 视觉压缩数据包含的信息有限，在不同任务上呈现出不同的分析效果。与分类任务相比，逐像素分割受到的损失更为严重。

4）联合训练结果：我们从三个维度评估联合训练计划的效果：压缩比、重建图像质量和机器感知性能。如 11）所示的雷达图所示，在极高压缩级别下，机器感知性能有所提升，而其他两个评估方面仅付出了有限的代价；同时

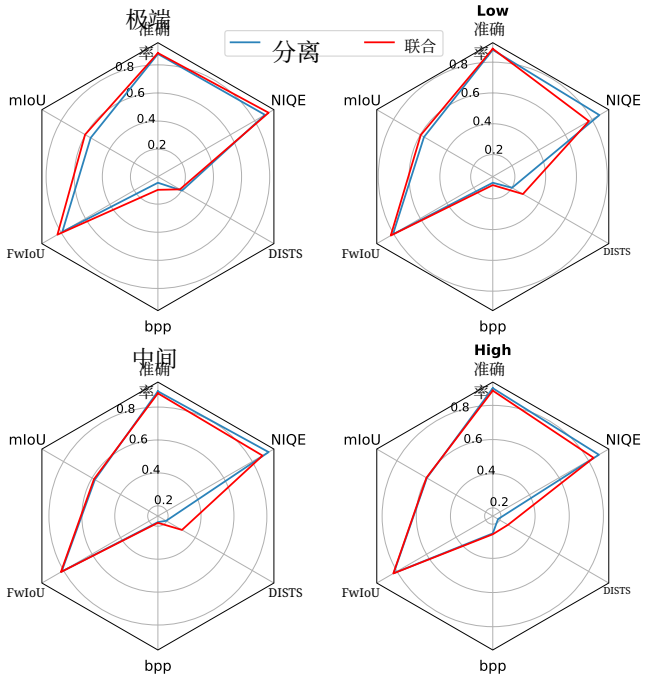

图 11. 联合训练计划与分离训练在压缩比、重建图像质量和机器感知性能方面的比较结果。结果来自我们所有压缩级别的模型。

在高压缩级别下，重建图像质量体现出不同质量评估之间的权衡。我们展示了机器感知-失真-人类感知权衡的存在，可将其视为对 [53] 中理论工作的现实实践拓展。

VI. 消融实验

A. 模型大小

我们从模型参数规模的角度说明所提出模型的可行性。我们将压缩数据输入分析模型与以原始 RGB图像为输入的代表性人脸属性预测模型进行比较。

如表 IV 所示，如果直接将压缩数据作为下游任务的输入，接收端可节省 50% 的存储成本。我们证明，与其他以 RGB图像为输入的方法相比，我们的方法能够减少存储占用，同时节省解码时间。在压缩视觉数据上直接进行分析，相较于对重建图像进行分析，可减少 3% 的参数量。

B. 编解码器超参数

超参数（通常记作 λ）是为实现图像压缩中失真与比特率（每像素比特数）之间不同的权衡而任意选取的。为探究式 (8) 中的超参数如何影响率失真权衡以及机器视觉任务的结果，我们在 λ_{MAE} 和 λ_{SSIM} 上进行了消融实验。表 V 和表 VI 展示了 λ_{MAE} 和 λ_{SSIM} 对 DISTS 和

表 IV

不同人脸属性预测方法的模型参数规模比较。参数量表示模型参数规模。对于我们的压缩数据输入分析方法，我们还将编码器模型的规模纳入考量，以模拟真实使用场景。我们方法的结果来自低比特率压缩比模型。

方法	编码器	解码器	准确率	参数量
单任务	✓	×	87.74%	26.9M
多任务	✓	×	88.01%	54.8M
多任务	✓	✓	88.12%	56.6M
刘 [54]			87.00%	100M
PSE [55]	×	-	91.23%	62M
DMTL [56]			92.10%	65M

表 V λ_{MAE}的消融研究。当 $\lambda_{MAE} = 10$时可实现最佳性能。

λ_{MAE}	DISTS↓	分类（%）
100	0.230	87.01
10	**0.201**	**87.76**
1	0.211	87.50

分类准确率。可以观察到，在 $\lambda_{MAE} = 10$, $\lambda_{SSIM} = 1$ 时，重建与分析的组合效果最佳。

C. 多任务分析模型损失

如损失函数公式（13）所示，通过超参数平衡各任务之间的关系。我们按照独立训练计划验证这两个任务之间的关系，并将分类权重 λ_{cls} 固定为 1.0，再调整分割任务的权重 λ_{seg}，以观察其性能变化。如图 12 所示，分类任务与分割任务之间存在权衡关系；当 $\lambda_{seg} = 2$ 时，视觉分析性能达到最佳。

D. 生物特征信息保留

为了评估重建图像在生物特征方面的准确性，我们采用两种图像处理任务：人脸特征检测和 1:1人脸识别。人脸关键点检测任务提供有关人脸结构变化的信息，而 1:1人脸识别任务则用于判断两张人脸图像是否属于同一个人。

人脸关键点检测：图像压缩应生成具有结构保真度和语义保真度的重建图像，以满足实际应用需求。考虑到人脸图像的特点，我们通过人脸关键点检测来评估结构保真度表现。我们将使用[58] 中的预训练检测模型对未压缩图像进行检测，并将检测结果设定为真实关键点 g_i。这里，归一化误差定义为：

$$Error = \frac{\|d_i - g_i\|_2}{d_{inter}} \tag{16}$$

超参数 λ_{SSIM} 的消融研究。 $\lambda_{SSIM} = 1$ 时性能最佳。

λ_{ssim}	DISTS↓	分类（%）
1.5	0.206	87.70
1	**0.201**	**87.76**
0.75	0.203	87.71

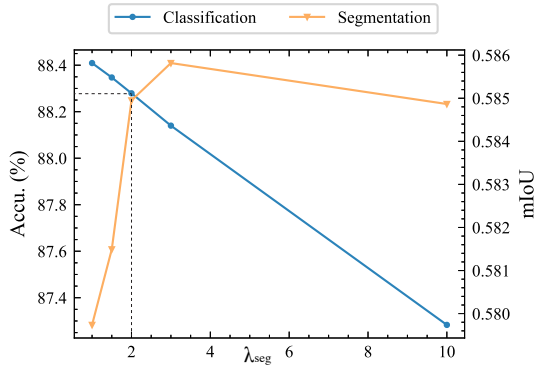

图 12. 调整超参数 λ_{seg} 后分类和分割任务的性能。交点具有 $\lambda_{seg} = 2$。

其中， d_i 是地标点 i 的预测二维坐标，d_{inter} 表示眼间距 [57]。*MeanError* 是所有地标点在公式（16）中 $Error$ 的平均值。

如图14所示，与 JPEG 相比，我们的压缩模型在高压缩比（即 ≤ 0.1每像素比特数）下具有更小的检测误差，并获得了保留更多结构信息的重建图像。

1:1人脸识别：我们计算原始图像与压缩图像之间的人脸识别 1:1 相似度。相似度越高，表示生物特征的损失越小。我们使用科大讯飞提供的在线 1:1人脸识别软件开发工具包[4]。比较结果如表 VII 所示：在极低码率约束下，以 JPEG编解码器作为比较方法时，所提出的方法可节省 67.24% 的比特率，同时在重建图像上实现可比的 1:1人脸识别性能。显然，JPEG 会产生严重的块效应和色偏，从而损害人脸图像中的生物特征信息，降低 1:1人脸识别中的相似度比较准确率。

E. 层次化表示效果

层次化表示有助于提升模型对视觉分析质量的鲁棒性。针对如何评估层次化表示的有效性这一问题，我们采用两种方法：1）定量分析，通过开展消融测试，确定层次化表示对视觉分析结果的影响。2）定性分析，确定潜在特征经过层次化分解后的视觉效果。

[4]https://global.iflytek.com/

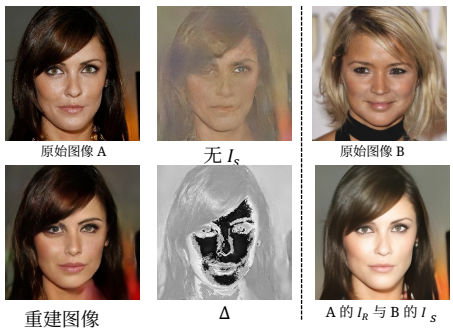

图 13. 当重建模型仅以内容通道表示作为输入，同时将风格通道表示全部置零进行屏蔽时的可视化结果。 Δ 表示重建图像 与无 I_S之间的抽象结果。右下角的图像是图像 A 的 I_R 特征与图像 B 的 I_S 特征组合而成的图像。

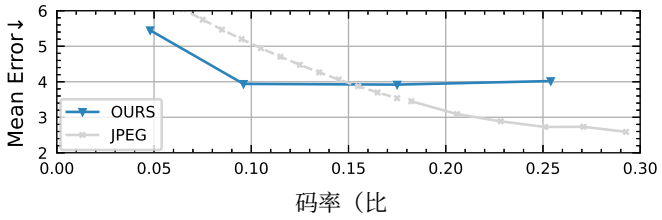

图14. 通过相对于眼间距的归一化误差 [57]表示的关键点检测误差。虚线表示极低比特率趋势下 JPEG 的曲线拟合。

定量分析：我们通过移除 I_S 特征，同时保持其余部分不变，来评估分层特征的效果。我们发现，在比特率相近的情况下，与采用非分层结构（即仅使用 I_R）相比，采用分层结构（即同时使用 I_R 和 I_S）能够使下游视觉分析任务的性能平均提升 1.66%。

定性分析： 如更新后的图 13 所示，我们从视觉角度分析所提出的层次化表示中各部分的作用。我们通过将语义分量 I_S 置零（如图像 A 所示），直接展示重建表示 I_R 的效果。 δ 图像是重建图像 A 与无 I_S之差，从视觉效果上可以直观地看出，语义分量 I_S 控制着亮度和颜色。如第 IV-A 节所定义， I_S 是一维特征，因此直接将其可视化并不能反映太多信息。于是，我们采用一种方法，从图像 A 中提取 I_R ，并从图像 B 中提取 I_S。随后将二者组合，通过组合图像观察I_S 的效果。从图 13 右下角的图像可以看出，面部和头发区域的亮度与颜色由目标图像 B 控制。

VII. 结论

本文提出了一种高效的分层生成式压缩模型，能够在压缩后的视觉

表 VII JPEG 与所提出压缩方法的重建图像上的 1:1人脸识别结果。我们将原始图像设为锚点，并使用它对重建图像进行 1:1人脸识别测试。我们从 CelebA-HQ 数据集的测试集划分中随机选取 200 张图像。通过率定义为 $\sum T(X,\hat{X}) > \xi \times 100\%$，其中阈值 ξ设为 0。67。

编解码器	码率（比特/像素）	通过率（%）
JPEG（qp=3）	0.182	64.0
JPEG（qp=8）	0.293	99.5
所提方法（提取）	0.096	99.5

以数据作为输入，同时，采用四种感知质量指标证明重建图像具有人类视觉保真度。本文开发了一种以压缩数据为输入的任务无关感知模型，以实现分析效率和灵活性，并高效支持多种分析任务。我们在人脸图像上进行了全面实验，验证了多种相互独立的视觉分析任务可以直接从压缩表示中进行推理，同时省去了解码过程和解码器端计算复杂度，体现了所提模型的实用价值。

本文的分析与观察为在潜在表示上执行视觉任务（尤其是基于学习得到的潜在表示）提供了可行方向。尽管基于学习的压缩模型在像素保真度指标（例如峰值信噪比和均方误差）方面存在局限性，但我们认为，基于学习的图像压缩模型的主要优势在于感知保留，以及为机器感知提供灵活的分析支持，而非保持信号级保真度；这为未来编解码器的发展指明了方向。

VIII. 参考文献

参考文献

[1] W. Yang、H. Huang、Y. Hu、L.-Y. Duan 和 J. Liu，《面向机器的视频编码：用于智能协同分析的紧凑视觉表示压缩》，ArXiv，第 abs/2110.09241 卷，2021年。[2] L.-Y. Duan、V. Chandrasekhar、J. Chen、J. Lin、Z. Wang、T. Huang、B. Girod 和 W. Gao，《MPEG-CDVS标准概述》，IEEE图像处理汇刊（TIP），第25卷，194，2016年1，179号，第 – 页。[3] L.-Y. Duan、Y. Lou、Y. Bai、T. Huang、W. Gao、V. Chandrasekhar、J. Lin、S. Wang 和 A. C. Kot，《视频分析紧凑描述符：新兴的MPEG标准》，IEEE多媒体，第26卷，54，2019年2，44号，第 – 页。[4] J. Ascenso、P. Akyazi、F. Pereira 和 T. Ebrahimi，《基于学习的图像编码：早期解决方案综述与主观质量评估》，载于《成像应用光学、光子学与数字技术VI》，P. Schelkens 和 T. Kozacki 编，第11353卷，国际光学与光子学学会。SPIE，2020年，第164 – 176 [ ] 10 1117 12 2555368 页。在线。可获[5]取地址：https://doi.org/ 。/ 。L.-Y. Duan、V. Chandrasekhar、J. Chen、J. Lin、Z. Wang、T. Huang、B. Girod 和 W. Gao，《MPEG-CDVS标准概述》，IEEE图像处理汇刊（TIP），第25卷，194，2015年1，179号，第 – 页。[6] S. R. Alvar 和 I. V. Baji´c，《面向协同智能的可压缩特征多任务学习》，载于IEEE国际图像处理会议（ICIP）。IEEE，2019年，第1705–1709页。[7] L. Gueguen、A. Sergeev、B. Kadlec、R. Liu 和 J. Yosinski，《直接从JPEG实现更快的神经网络》，《神经信息处理系统进展》（NIPS），第31卷，第3933–3944页，2018年。[8] M. Ehrlich 和 L. S. Davis，《JPEG变换域中的深度残差学习》，载于IEEE/CVF国际计算机视觉会议（CVPR）论文集，2019年，第3484–3493页。

[9]R. Torfason、F. Mentzer、E. Agustsson、M. Tschannen、R. Timo- fte 和 L. Van Gool，《无需解码的深度压缩图像理解》，国际学习表征会议（ICLR），2018年。[10] K. Liu、D. Liu、L. Li、N. Yan 和 H. Li，《基于学习得到的可逆表征的语义到信号可扩展图像压缩》，国际计算机视觉期刊（IJCV），第1–17页，2021年。[11] J. Ball´e、V. Laparra 和 E. P. Simoncelli，《端到端优化图像压缩》，国际学习表征会议（ICLR），第1–27页，2017年。[12] Z. Cheng、H. Sun、M. Takeuchi 和 J. Katto，《采用离散高斯混合似然和注意力模块的学习式图像压缩》，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2020年，第7936–7945页。[13] J. Redmon、S. K. Divvala、R. B. Girshick 和 A. Farhadi，《你只需看一次：统一的实时目标检测》，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集。IEEE计算机学会，2016年，第779–788页。 [在线]。可访问：https://doi.org/10 1109 2016 91 。/CVPR。K. Simonyan 和 A. Zisserman，[14]《用于大规模图像识别的超深卷积网络》，载于国际学习表征会议（ICLR），2015年。[15] L.-C. Chen、G. Papandreou、I. Kokkinos、K. Murphy 和 A. L. Yuille，《Deeplab：基于深度卷积网络、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（TPAMI），第40卷，第4期，第834–848页，2017年。[16] M. Oquab、P. Stock、D. Haziza、T. Xu、P. Zhang、O. Celebi、Y. Has- son、P. Labatut、B. Bose-Kolanu、T. Peyronel 和 C. Couprie，《利用深度生成模型实现低带宽视频聊天压缩》，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR）研讨会论文集，2021年6月，第2388–2397页。[17] I. Qualcomm Technologies，《移动设备上的实时神经视频解码》，https://www.qualcomm.com/news/onq/2021/06/17/worlds-first-software-based-neural-video-decoder-running-hd-for mat-real-time，2021年。[18] D. Marr，《视觉：对人类表征的计算研究》，1983年。[19] X. Huang、M.-Y. Liu、S. Belongie 和 J. Kautz，《多模态无监督图像到图像转换》，载于欧洲计算机视觉会议（ECCV），2018年。[20] H.-Y. Lee、H.-Y. Tseng、Q. Mao、J.-B. Huang、Y.-D. Lu、M. Singh 和 M.-H. Yang，《Drit++：通过解耦表征实现多样化的图像到图像转换》，国际计算机视觉期刊（IJCV），第128卷，第10期，第2402–2417页，2020年。[21] Y. Choi、M. Choi、M. Kim、J.-W. Ha、S. Kim 和 J. Choo，《StarGAN：用于多域图像到图像转换的统一生成对抗网络》，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2018年。[22] P. Isola、J.-Y. Zhu、T. Zhou 和 A. A. Efros，《利用条件对抗网络实现图像到图像转换》，2017年 IEEE计算机视觉与模式识别会议（CVPR），第5967–5976页，2016年。[23] J. Chang、Q. Mao、Z. Zhao、S. Wang、S. Wang、H. Zhu 和 S. Ma，《通过深度语义合成实现分层概念图像压缩》，载于 IEEE国际图像处理会议（ICIP），2019年，第694–698页。[24] J. Chang、Z. Zhao、L. Yang、C. Jia、J. Zhang 和 S. Ma，《千比一：概念编码的语义先验建模》。IEEE，2021年，第1–6页。[25] M. Heusel、H. Ramsauer、T. Unterthiner、B. Nessler 和 S. Hochreiter，《通过双时间尺度更新规则训练的生成对抗网络收敛于局部纳什均衡》，神经信息处理系统进展（NIPS），第6626–6637页，2017年。[26] K. Ding、K. Ma、S. Wang 和 E. P. Simoncelli，《用于优化图像处理系统的图像质量模型比较》，国际计算机视觉期刊（IJCV），第1258–1281页，2021年。[27] R. Zhang、P. Isola、A. A. Efros、E. Shechtman 和 O. Wang，《深度特征作为感知度量的惊人有效性》，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2018年，第586–595页。[28] G. K. Wallace，《JPEG静态图像压缩标准》，IEEE消费电子汇刊，第38卷，第1期，第xviii–xxxiv页，1992年。[29]B. Bross、Y. Wang、Y. Ye、S. Liu、J. Chen、G. J. Sullivan 和 J. Ohm，《多功能视频编码（VVC）标准及其应用概述》，IEEE电路与系统汇刊：视频。 3264, 2021。[在线]。可访问：https://doi.org/10 1109 2021 3101953 。/TCSVT。S. F. d. [30]Santos 和 J. Almeida，“少即是多：直接从 JPEG 加速更快的神经网络”，ArXiv 预印本 arXiv:2104.00185，2021。[31] A. Chadha、A. Abbas 和 Y. Andreopoulos，“ 使用卷积神经网络进行视频分类：将编解码器用作时空活动传感器”，IEEE视频技术电路与系统汇刊（TCSVT），第 29 卷，第 2 期，第 475–485 页，2017。[32] C.-Y. Wu、M. Zaheer、H. Hu、R. Manmatha、A. J. Smola 和 P. Kr¨ahenb¨uhl，“压缩视频动作识别”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2018，第6026–6035 页。[33] J. Ball´e、D. Minnen、S. Singh、S. J. Hwang 和 N. Johnston，“基于尺度超先验的变分图像压缩”，国际学习表征会议（ICLR），第 1–23 页，2018。[34] E. Agustsson、F. Mentzer、M. Tschannen、L. Cavigelli、R. Timofte、L. Benini 和 L. Van Gool，“用于端到端学习可压缩表征的软到硬矢量量化”，神经信息处理系统进展（NIPS），2017。[35] Y. Hou、L. Zheng 和 S. Gould，“学习用少量颜色构造图像”，载于 IEEE/CVF计算机视觉与模式识别会议（CVPR），2020，第 10 113–10 122 页。[36] J. Choi 和 B. Han，“用于 JPEG图像压缩的任务感知量化网络”，载于欧洲计算机视觉会议（ECCV），A. Vedaldi、H. Bischof、T. Brox 和 J.-M. Frahm 编，Springer International Publishing，2020，第 309–324 页。[37] W. Shurun、W. Zhao、W. Shiqi 和 Y. Yan，“面向机器视觉的端到端压缩：网络架构设计与优化”，IEEE开放电路与系统期刊（OJCAS），2021。[38] L. Duan、J. Liu、W. Yang、T. Huang 和 W. Gao，“面向机器的视频编码：协同压缩与智能分析范式”，IEEE图像处理汇刊（TIP），第 29 卷，第8680–8695 页，2020。[39] Z. Huang、C. Jia、S. Wang 和 S. Ma，“HMFVC：一种人机友好型视频压缩方案”，IEEE视频技术电路与系统汇刊（TCSVT），2022。[40] F. Codevilla、J. G. Simard、R. Goroshin 和 C. Pal，“用于机器感知的学习式图像压缩”，2021。[41] M. Akbari、J. Liang 和 J. Han，“DSSLIC：基于深度语义分割的分层图像压缩”，载于 IEEE声学、语音与信号处理国际会议（ICASSP），2019，第 2042–2046 页。[42] Y. Hu、S. Yang、W. Yang、L.-Y. Duan 和 J. Liu，“面向人类与机器视觉编码：一种可扩展图像编码方法”，载于 IEEE多媒体与博览会国际会议（ICME）。IEEE，2020，第 1–6 页。[43] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2016，第 770–778 页。[44] F. Mentzer、G. D. Toderici、M. Tschannen 和 E. Agustsson，“高保真生成式图像压缩”，神经信息处理系统进展（NIPS），第 33 卷，2020。[45] T.-C. Wang、M.-Y. Liu、J.-Y. Zhu、A. Tao、J. Kautz 和 B. Catanzaro，“利用条件生成对抗网络进行高分辨率图像合成与语义操控”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2018。[46] O. Sener 和 V. Koltun，“作为多目标优化的多任务学习”，载于神经信息处理系统进展（NIPS），S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编。Curran Associates, Inc.，2018，第 525–536 页。 [在线]。可访问：http://papers.nips.cc/paper/7334-multi-task-learning-as-multi-objective-optimization.pdf[47] C.-H. Lee、Z. Liu、L. Wu 和 P. Luo，“MaskGAN：实现多样化且交互式的人脸图像操控”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2020。[48] Z. Liu、P. Luo、X. Wang 和 X. Tang，“野外大脸属性深度学习”，载于 IEEE国际计算机视觉会议（ICCV）论文集，2015 年 12 月。[49] J. Deng、W. Dong、R. Socher、L.-J. Li、K. Li 和 L. Fei-Fei，“ImageNet：大型分层图像数据库”，载于 2009 IEEE计算机视觉与模式识别会议，2009，第 248–255 页。[50] A. Mittal、R. Soundararajan 和 A. C. Bovik，“构建一种‘完全盲’图像质量分析器”，IEEE信号处理快报，第 20 卷，第 3 期，第 209–212 页，2012。[51] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，国际学习表征会议（ICLR），2015。

[52] X. Mao、Q. Li、H. Xie、R. Y. Lau、Z. Wang 和 S. Paul S molley，“最小二乘生成对抗网络”，载于 IEEE国际计算机视觉会议（ICCV）论文集，2017，第 2794–2802 页。[53] D. Liu、H. Zhang 和 Z. Xiong，“论分类—失真—感知权衡”，神经信息处理系统进展（NIPS），2019。[54] Z. Liu、P. Luo、X. Wang 和 X. Tang，“野外人脸属性深度学习”，载于 IEEE国际计算机视觉会议（ICCV）论文集，2015，第 3730–3738 页。[55] Y. Zhang、W. Shen、L. Sun 和 Q. Li，“用于面部属性分析的位置压缩与激励模块”，载于英国机器视觉会议（BMVC）。BMVC，2018。[56] H. Han、A. K. Jain、F. Wang、S. Shan 和 X. Chen，“异构人脸属性估计：一种深度多任务学习方法”，IEEE模式分析与机器智能汇刊（TPAMI），第 40 卷，第 11 期，第 2597–2609 页，2017。[57] O. Celiktutan、S. Ulukaya 和 B. Sankur，“人脸标志点定位技术的比较研究”，EURASIP图像与视频处理期刊（JIVP），第 2013 卷，第 1 期，第 1–27 页，2013。[58] V. Kazemi 和 J. Sullivan，“利用回归树集成实现一毫秒人脸对齐”，载于 IEEE计算机视觉与模式识别会议（CVPR）论文集，2014，第 1867–1874 页。