arXiv:2303.10372v1 [cs.CV] 18 Mar 2023

恰可察觉　的视觉冗余预测：一种深度多模态驱动方法

Wuyuan Xie、Shukang Wang、Sukun Tian、Lirong Huang、Ye Liu、Miaohui Wang *

摘要

恰可察觉差异（JND）是指人眼无法察觉的最大视觉变化，在多媒体系统中有广泛的应用。然而，现有的大多数 JND 方法只关注单一模态，很少考虑多模态信息的互补作用。本文从端到端同源多模态视角研究刚可觉差建模，提出 hmJND-Net。具体而言，我们探索了三种重要的视觉敏感模态，包括显著性、深度和分割。为更好地利用同源多模态信息，我们通过求和增强和减法偏移建立了一种有效的融合方法，并基于自注意力驱动的编码器-解码器范式对齐同源多模态特征。在八个不同基准数据集上的大量实验结果验证了 hmJND-Net 优于八种具有代表性的方法。

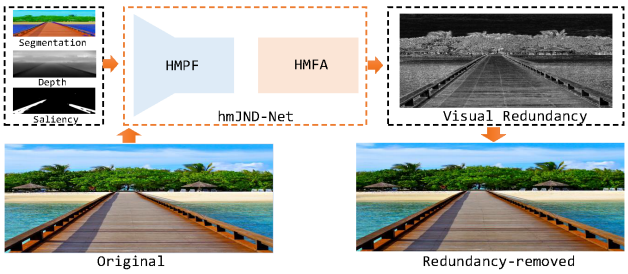


图1：**所提出的多模态驱动视觉冗余预测网络示意图。**该网络基于同源模态先验融合（HMPF）模块和同源模态特征对齐（HMFA）模块，联合考察显著性、深度、分割和RGB模态，以预测视觉冗余。

引言

恰可察觉差异（JND）反映了人类视觉系统（HVS）无法察觉的最大视觉冗余（也称为可见性阈值</style）。在多媒体领域，可见性阈值能有效量化视觉信号中的感知冗余，并广泛应用于图像压缩和处理系统（ Wang 等人，2022）。例如，JND 可指导多媒体压缩中冗余信息的去除（Zhang et al. 2016; Tian 等人，2021），在不影响感知质量的前提下提高编码效率。它还可提高质量评估的准确性（Seo, Ki, and Kim 2020），并增强水印嵌入的强度（Liu et al. 2021a）。根据建模方法，现有的 JND 方法大致可分为：1）人类视觉系统引导模型（Bae, Kim, and Kim 2016; Wu et al. 2017; Chen and Wu 2019）和 2）基于学习的模型（Huang 等人，2017；Ki 等人，2018；Liu 等人，2019；Shen 等人，2020；Tian 等人，2021）。

人类视觉系统引导模型主要在像素域或变换域中探索可见性阈值，这两者遵循相同的基本流程（Liu、Zhan 和 Wang 2020）：首先对若干显著的视觉掩蔽效应进行建模，然后通过非线性融合方法将它们结合起来。根据人类视觉系统的特性，现有方法考虑的影响因素包括亮度适应（LA）、空间对比度掩蔽（对比度掩蔽）、视觉显著性（Wang et al. 2014）、对比度敏感度（Bae、Kim 和 Kim 2016）、图案复杂度（Wu et al. 2017）和中央凹效应（Chen and Wu 2019），等。然而，人类视觉系统引导方法至少存在两个缺点：1）由于我们对人类视觉系统的认知有限，无法获取预测恰可察觉视觉冗余所需的所有可能影响因素；2）现有解析模型无法准确刻画不同人类视觉系统因素之间潜在的相互关系。

随着深度学习的发展，一些初步研究（Liu 等，2019；Shen 等人，2020；Tian 等人，2021）旨在解决上述人类视觉系统引导模型存在的问题。研究人员构建了若干基于编码器的恰可察觉差异数据集，这些数据集对于数据驱动方法至关重要。在标注可见性阈值图时，每位受试者都会评估一系列由典型编解码器采用不同质量因子生成的失真图像或视频。例如，

*本研究部分得到了国家自然科学基金（项目编号：61701310和61902251）、深圳市自然科学基金（项目编号：20200805200145001和JCYJ20220809160139001）以及广东省自然科学基金（项目编号：2021A1515011877和2022A1515011245）的资助。（通讯作者：*Miaohui Wang*）
版权所有 © 2023，人工智能促进协会（www.aaai.org）。保留所有权利。

（Jin 等，2016）构建了一个基于 JPEG 压缩的数据集 MCL-JCI：每幅源图像都由 JPEG 编解码器以 1 到 100 的质量因子编码 100 次，随后志愿者通过二分搜索策略评估解码后的图像，从而确定逐图像的标签数据。（Shen 等人，2020）进一步构建了一个基于多功能视频编码（VVC）编解码器的大型数据集。此外，（Wang 等，2016）构建了一个逐视频数据集。然而，基于学习的 JND 模型严重依赖人工标注的数据集，并面临数据不足的问题。鉴于此，我们尝试通过引入同源多模态来探索训练数据，从而进一步提高视觉冗余的预测准确度。

受上述讨论的启发，我们研究了一种端到端的同源多模态驱动视觉冗余预测网络，如图 1 所示。该网络融合了显著性、深度、分割和 RGB 图像模态。具体而言，我们首先从输入数据中获取深度、显著性和分割模态。随后，通过通道注意力网络融合这些先验多模态特征，再利用 Transformer 网络将 RGB 模态与融合后的先验多模态特征对齐，从而获得最终的可见性阈值。

与先前方法相比，本文的主要贡献总结如下。我们首次在监督式可见性阈值建模中研究同源多模态信息，并设计了一种新颖的端到端同源多模态 JND 预测网络（即 hmJND-Net）。所提出的 hmJND-Net 缓解了训练数据不足的问题，同时提高了预测准确率。为更好地利用深度、显著性和分割这三种先验模态，我们进一步开发了一种新的同源模态先验融合（HMPF）模块，以提前融合这些先验特征；此外，我们还设计了一种采用 Transformer 结构的同源模态特征对齐（HMFA）模块，用于对齐 RGB 图像模态与三种先验模态。在八个基准数据集上的实验结果表明，与具有代表性的方案相比，我们的 hmJND-Net 能够提供更高的准确率和更多的比特率节省。

相关工作

本节首先简要回顾一些具有代表性的恰可察觉视觉冗余模型，包括人类视觉系统引导的方法和基于学习的方法。随后，我们介绍提出 hmJND-Net 的动机。

人类视觉系统引导方法

如前所述，人类视觉系统引导模型的基础源自人类视觉系统的特性。在像素域中，亮度适应和对比度掩蔽效应被广泛采用：亮度适应效应反映了人类视觉系统对具有不同背景亮度的视觉信号的敏感度差异，而对比度掩蔽效应则反映了人类视觉系统对均匀或非均匀内容的微小变化（即可见性抑制）的响应。例如，(Chou 和 Li，1995) 提出了一种经典的 JND 框架，将亮度适应建模为准抛物线模型，并根据局部梯度强度推导出对比度掩蔽。在此基础上，



图2：**深度模态和分割模态示例**。

(Yang 等，2005) 开发了掩蔽非线性加和模型（NAMM）模块，以减弱两种掩蔽效应之间的叠加效应。(Liu 等，2010) 进一步将原始图像分解为结构分量和纹理分量，以估计边缘掩蔽和纹理掩蔽，旨在避免高估边缘区域和纹理区域。(Wu 等，2013) 基于自由能原理建立了自回归模型，用于预测有序区域和无序区域，并在可见性阈值估计中采用无序掩蔽效应。随后，他们利用方向的多样性来衡量图案复杂度，并推导出图案掩蔽 (Wu 等，2017)。最近，(Chen and Wu, 2019) 进一步将中央凹效应和非对称视觉敏感度引入最大视觉冗余建模。(Jiang et al.，2022) 探索了 Karhunen–Lo`eve 变换（KLT）系数的特性，以推导出感知无损预测。

除了像素域模型外，变换域方法也已出现，以促进多媒体应用。对比度敏感度函数（CSF）在变换域刚可觉差建模中起主导作用，它反映了人类视觉系统的频率特性。例如，(Peterson and Ahumada Jr 1992) 将 CSF 分量建模为空间对比度的指数函数。(Zhang, Lin, and Xue 2005) 基于离散余弦变换（DCT）域中的块分类，修订了亮度适应和对比度掩蔽模型。(Wei and Ngan 2009) 使用伽马校正来补偿 DCT 域中的亮度适应和对比度掩蔽效应。(Bae, Kim, and Kim 2016) 基于不同尺寸 DCT 块的结构对比度建立了对比度掩蔽效应，设计了一种基于 DCT 的局部冗余概率来估计变换系数的分布，并将其引入感知视频编码。

基于学习的方法

随着深度学习的发展，基于学习的刚可觉差估计已成为该领域的新趋势。例如，(Huang et al. 2017) 提出基于提取的视频特征，学习基于组的可见性阈值分布的均值。(Ki et al. 2018) 建立了一种基于能量降低的模型，采用线性回归和基于卷积神经网络（CNN）的模型来减轻压缩伪影的影响。(Liu et al. 2019) 开发了一种深度二分类器，用于预测编码图像相对于其

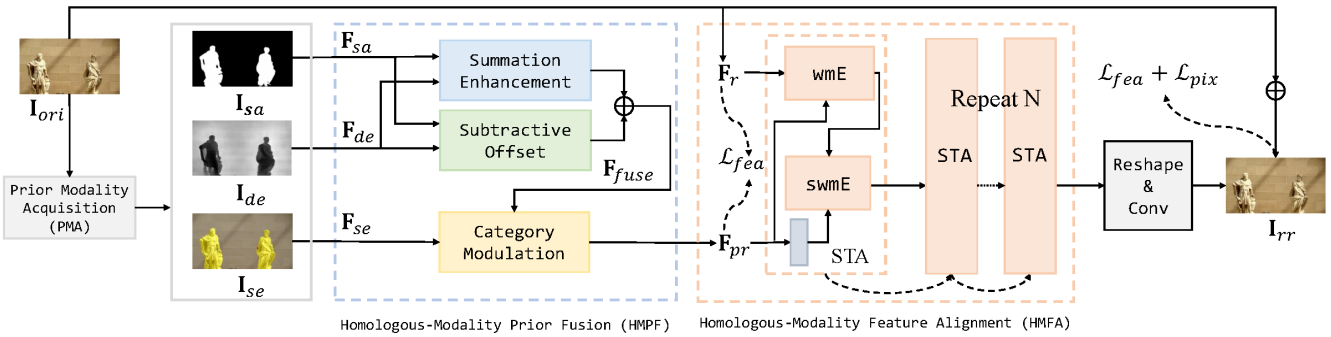


图3：**所提出的基于多模态的视觉冗余框架流程**。该框架主要由先验模态获取（PMA）模块、同源多模态先验融合（HMPF）模块和同源多模态特征对齐（HMFA）模块组成。**请在电子版中放大查看，以便了解更多细节**。

原始图像是否存在感知失真。(Shen et al. 2020) 将输入图像分解为三个分量（即亮度、对比度和结构），然后训练深度 CNN 退化模型来预测可见性阈值。(Tian et al. 2021) 设计了一种块级视觉冗余算法，利用 AlexNet 预测每个图像块的编码参数。(Wang et al. 2022) 通过考虑斜向校正效应来预测可见性阈值，并最终建立了可见性阈值与质量因子之间的端到端 CNN 映射。

动机

近年来，多模态数据通常用于从不同来源或不同视角描述物体。来自不同来源的数据称为异构多模态信息（Yu 等人，2021；Min 等人，2020）。来自单一来源的数据称为同源多模态信息（Wang 等人，2018；Shao 和 Wang，2021；Xie、Huang 和 Wang，2022）。如何融合同源多模态数据是多媒体图像处理领域的热门研究课题。许多研究利用同源多模态信息来提升图像处理系统的性能。例如，Wang 等人（2018）引入分割模态，以解决超分辨率纹理不真实的问题。Cheng、Fu 和 Yang（2020）采用深度模态提取高分辨率和低分辨率图像块，用于训练超分辨率网络。Shao 和 Wang（2021）在图像修复任务中利用显著性模态作为细节重建的监督信息。

受上述工作的启发，本文首次研究基于同源多模态的 JND 建模（即显著性、深度和分割）：1）**显著性模态**。由于人类视网膜的视敏度会随着偏离中央凹的距离增加而降低，可见性阈值会随偏心度增加。因此，自然应对显著区域减少较少的视觉冗余，对非显著区域减少较多的视觉冗余。2）**深度模态**。靠近相机的图像块会更清晰，而深度越深，图像内容就会越小、越模糊。如图2所示，视觉冗余与其深度有关，深度较深的区域比深度较浅的区域能容忍更多视觉冗余。基于这一先验知识，我们可以基于深度模态，在更深的深度区域减少更多视觉冗余。3）**分割模态**。我们发现，不同物体具有不同的可见性阈值。如图2所示，具有复杂纹理的树木和房屋，其可见性阈值显著高于平坦天空和墙壁。因此，我们通过语义分割模态将类别信息引入框架，从而针对不同物体减少不同程度的冗余。

提出的多模态框架

本节详细介绍所提出的端到端同源多模态可见性阈值预测网络，该网络由先验模态获取（PMA）模块、同源模态先验融合（HMPF）模块和同源模态特征对齐（HMFA）模块组成，如图3所示。

概述

问题表述。与现有的可见性阈值方法不同，我们以端到端的方式构建逐像素同源多模态网络，因此，我们的总体任务可以表示为通过最小化一个专门设计的优化问题：

$$\min_{I_{rr}} \left(\mathcal{L}_{overall}\left(\mathcal{L}_{fea}(F_r, F_{pr}), \mathcal{L}_{pix}(I_{rr}, I_{gt})\right)\right), \quad (1)$$

其中，$\mathcal{L}_{overall}(\cdot)$ 表示总体损失函数。$\mathcal{L}_{overall}(\cdot)$ 通过对 RGB 特征 F_r 与先验融合特征 F_{pr} 之间的相似度度量$\mathcal{L}_{fea}$，以及输出去冗余图像 I_{rr} 与其真实值 I_{gt}之间的像素距离 $\mathcal{L}_{pix}$ 进行加权得到。

具体而言，

$$\begin{aligned}&\mathcal{L}_{overall}\left(\mathcal{L}_{fea}(F_r, F_{pr}), \mathcal{L}_{pix}(I_{rr}, I_{gt})\right)\\&= \lambda_{fea} \times \mathcal{L}_{fea}(F_r, F_{pr}) + \lambda_{pix} \times \mathcal{L}_{pix}(I_{rr}, I_{gt})\\&= \frac{\lambda_{fea}}{h_f \times w_f \times c_f}\|F_r - F_{pr}\|_1 + \frac{\lambda_{pix}}{h_p \times w_p \times c_p}\|I_{rr} - I_{gt}\|_2^2\end{aligned}, \quad (2)$$

其中 (h_f, w_f, c_f) 表示先验融合特征的高度、宽度和通道数。(h_p, w_p, c_p) 表示

预测值。 $\mathcal{L}_{fea}$ 表示RGB特征与先验融合特征之间的逐像素 L_1 损失，用于衡量二者的相似度，并通过最小化二者之间的距离实现多模态特征对齐。 $\mathcal{L}_{pix}$ 表示输出去冗余图像与真实值之间的逐像素 L_2 损失。 λ_{fea} 和 λ_{pix} 是相应的权重。

架构。 我们在所提出的 hmJND-Net 中研究了四种模态，包括 RGB图像模态、显著性模态、深度模态和分割模态。图3展示了整体网络架构，其形式化表示为

$$I_{rr} = \mathcal{M}_{JND}\left(I_{ori}, \mathcal{M}_{EX}(I_{hm})\right), \tag{3}$$

其中， I_{hm} 表示通过原始RGB模态 I_{ori}获取的同源多模态信息，包括显著性模态 I_{sa}、深度模态 I_{de}和分割模态 I_{se}。通常，式（3）可分解为两个子任务：多模态特征提取模块$\mathcal{M}_{EX}$ 和多模态预测模块 $\mathcal{M}_{JND}$。

多模态特征提取模块 $\mathcal{M}_{EX}$由 PMA 和 HMPF 模块组成。首先，将原始 RGB 图像模态输入 PMA 模块，以生成相应的三种先验模态，并通过三层卷积提取浅层特征。随后，将浅层特征输入 HMPF 模块 $\mathcal{M}_{HMPF}$，以融合先验模态特征。 $\mathcal{M}_{EX}$（·）可表示为

$$F_{pr} = \mathcal{M}_{HMPF}\left(\mathcal{M}_{PMA}(I_{ori})\right), \tag{4}$$

其中， F_{pr} 表示经过同源多模态融合模块处理后的先验融合特征，$\mathcal{M}_{PMA}$(·)表示一个 PMA 模块， $\mathcal{M}_{HMPF}$(·)表示一个 HMPF 模块。

经过 HMPF 处理后， F_{pr} 和 F_r 一起输入 HMFA 模块进行特征对齐。首先，最小化 F_{pr} 与 F_r 之间的逐像素距离，然后使用若干个 Swin Transformer对齐（STA）块进一步对齐 F_{pr} 和 F_r 特征。随后，将对齐后的特征输入重塑模块。最后，将初始输入连接起来进行残差学习。$\mathcal{M}_{JND}$(·)可表示为

$$I_{rr} = \theta\left(\mathcal{M}_{RE}(\mathcal{M}_{HMFA}(F_{pr}, F_r))\right), \tag{5}$$

其中， F_r 表示通过三层卷积提取的 RGB图像模态中间特征，$\mathcal{M}_{HMFA}$(·)表示一个 HMFA 模块， $\mathcal{M}_{RE}$(·)表示一个重塑模块， θ(·)表示归一化操作，用于将输出限制在合理范围内。

同源模态先验融合

为了充分利用先验多模态信息，我们设计了一个特殊的同源模态先验融合（HMPF）模块，如图 4 所示。在 HMPF 处理之前，使用三层 3×3 卷积提取获取的先验模态，以获得相应的浅层特征 F_x($i.e.$, $x = \{sa, de, se\}$)。换言之，I_{sa}、 I_{de} 和 I_{se} 将分别嵌入 F_{sa}、 F_{de} 和 F_{se}。鉴于深度和显著性模态在

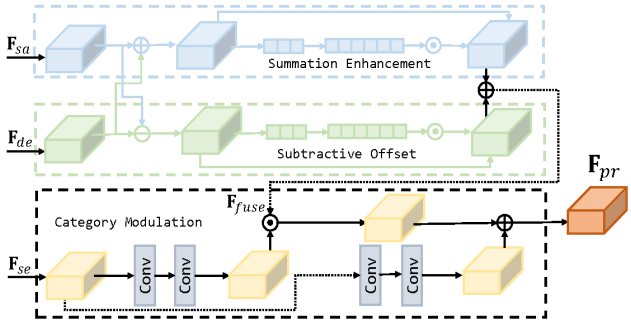


图 4：**同源模态先验融合（HMPF）**。输入是显著性$\mathbf{F}_{sa}$、深度$\mathbf{F}_{de}$ 和分割$\mathbf{F}_{se}$ 模态，输出是先验融合特征$\mathbf{F}_{pr}$。

可见性阈值估计中并非总是呈正相关，深度较大的图像区域可能属于显著区域，而深度较浅的区域则可能属于非显著区域。因此，为了捕捉同源多模态互补信息，hmJND-Net 不仅需要学习深度和显著性模态之间的一致性，还需要学习它们之间的差异。

基于上述分析，我们提出一种包含两个子模块的策略，即求和增强 $\mathcal{M}_{SE}$ 和减法偏移 $\mathcal{M}_{SO}$，以学习这两种模态之间的互补关系，然后通过引入类别信息来调制融合特征图 F_{fuse}。具体而言，我们通过将 F_{sa} 与F_{de}相减，得到包含差异信息的差减特征F_{sub}；并通过将 F_{sa} 与 F_{de}相加，得到包含一致性信息的求和特征 F_{sum}。

受通道注意力网络 SENet（Hu、Shen 和 Sun（2018））的启发，我们进一步对这两个中间特征进行通道注意力加权，以获得增强特征 F_e 和偏移特征 F_o，然后将它们融合得到 F_{fuse}，其公式如下：

$$F_{fuse} = F_e \oplus F_o, \tag{6}$$

$$F_e = F_{sum} \otimes \mathcal{F}^s_{fc}(\mathcal{F}^r_{fc}(\mathcal{F}_{gp}(F_{sum}))) \oplus F_{sum}, \tag{7}$$

$$F_o = F_{sub} \otimes \mathcal{F}^s_{fc}(\mathcal{F}^r_{fc}(\mathcal{F}_{gp}(F_{sub}))) \oplus F_{sub}, \tag{8}$$

其中， $\oplus$ 和 $\otimes$ 分别表示逐元素求和和逐元素相乘。 $\mathcal{F}_{gp}$(·)表示全局池化， $\mathcal{F}^x_{fc}$(·)表示全连接层，后接 Sigmoid 归一化或 ReLU 激活层。

获得显著性和深度模态的融合特征后，我们需要融入类别信息。受空间特征变换（Wang 等人，2018）的启发，我们提出使用语义分割模态，并以此调制融合特征。该融合过程可表示为

$$F_{pr} = F_{fuse} \otimes \mathcal{F}_{conv}(F_{se}) \oplus \mathcal{F}_{conv}(F_{se}), \tag{9}$$

其中， $\mathcal{F}_{conv}$(·)表示一个两层卷积。三种模态的融合特征将被送入下一个 HMFA 模块。

同源模态特征对齐

获得 F_{pr} 后，由于 RGB 模态特征 F_r 与先验多模态特征 F_{pr} 之间

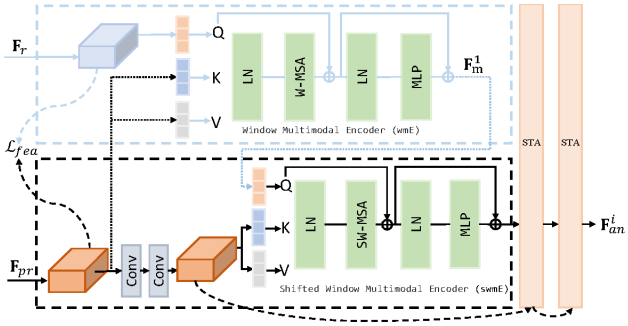


图 5：**同源模态特征对齐 (HMFA)**。输入为先验融合特征 $\mathbf{F}_{pr}$ 和 RGB 模态特征 $\mathbf{F}_r$，输出为对齐后的特征$\mathbf{F}_{an}^i$。

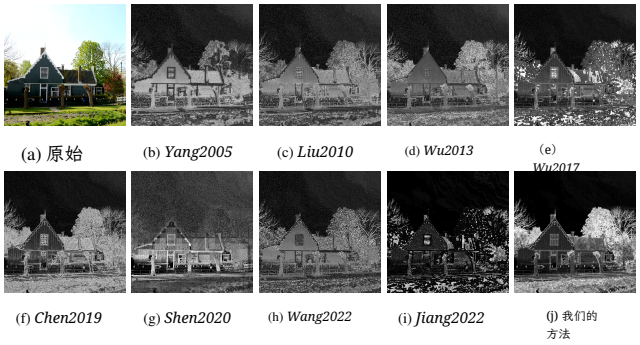


图6：**9种具有代表性方法预测的视觉冗余度对比**。在图(b)–(j)中，较亮的像素表示较高的冗余度。

存在较大差异，我们需要对二者进行对齐。为此，我们进一步设计了一个同源模态特征对齐（HMFA）模块，如图 5 所示。我们将这两个特征的相似度作为显式对齐，并通过受 Liu等人，2021b 启发的 Swin Transformer对齐（STA）模块进行隐式对齐。其主要目的是利用自注意力机制学习 F_r 和 F_{pr} 之间的对齐关系。

首先，HMFA 模块由多个 STA 模块组成，每个 STA 模块基于多头注意力机制，将 RGB 模态与先验同源多模态特征进行对齐。

$$F_{an} = \mathcal{B}_{STA}(F_{pr}, F_r), \tag{10}$$

其中，F_{an} 表示 STA 模块对齐后的输出特征。$\mathcal{B}_{STA}(\cdot)$ 表示一个 STA 模块，其中包含两个特征编码器：窗口多模态编码器 (wmE) 和移位窗口多模态编码器 (swmE)。STA 模块按照式 (11) 组织。

$$F_{an}^{i+1} = \begin{cases} \mathcal{B}_{sw}^i(Q = F_m^i, K, V = \mathcal{F}_{conv}^i(F_{pr})) & i = 1 \\ \mathcal{B}_{sw}^i(Q = F_m^i, K, V = \mathcal{F}_{conv}^{2i-1}(F_{pr})) & n \geq i > 1 \end{cases}, \tag{11}$$

其中，$\mathcal{B}_w^i(Q, K, V)$ 和 $\mathcal{B}_{sw}^i(Q, K, V)$ 分别表示第 i 个 wmE 和 swmE。(Q, K, V) 是自注意力机制中的查询、键和值。F_M^i 表示第i 个 wmE 的输出，其定义为

$$F_m^i = \begin{cases} \mathcal{B}_w^i(Q = F_r, K, V = F_{pr})) & i = 1 \\ \mathcal{B}_w^i(Q = F_{an}^{i-1}, K, V = \mathcal{F}_{conv}^{2i-2}(F_{pr}))) & n \geq i > 1 \end{cases}, \tag{12}$$

其中，$\mathcal{F}_{conv}^i(\cdot)$ 表示第i个卷积块，每个卷积块由两层卷积组成。

wmE 与 swmE 的区别主要在于用于不同窗口间交互的移位窗口机制；它们在 STA 模块中交替使用。这种结构的优势在于，通过自注意力机制反复连接先验多模态特征，增强深度网络的记忆能力，逐步捕获多模态特征的全局信息，学习同源多模态之间的相互关系，并对齐 RGB 与先验模态。

实验验证

实验方案

数据集描述。我们基于最新的基准数据集（Shen 等人，2020）训练了 hmJND-Net。该数据集涵盖多种图像内容，包括户外、室内、风景、自然场景、人物、物体和建筑物。数据集由 202 张尺寸为 1920×1080 的高清原始图像以及通过 VVC 生成的 7878 张去冗余图像组成。每张原始图像都有 39 个具有不同去冗余级别的编码版本，这些版本由主观实验选出。

实验中使用的三种模态均由原始图像生成，其中深度模态由DPT-Hybrid（Ranftl、Bochkovskiy和Koltun，2021）生成，显著性模态由BASNet（Qin等人，2019）生成，分割模态由BEIT（Bao等人，2021）生成。图7展示了一些典型示例。

我们按照**8:1:1**的比例从数据集中随机抽取图像对，分别作为训练集、验证集和测试集。为展示泛化能力，我们还在另外七个基准数据集上独立进行了测试实验，包括*CSIQ*（Larson和Chandler，2010）（30幅尺寸为 512×512的原始自然内容图像）、*KADID-10K*（Lin、Hosu和Saupe，2019）（81幅尺寸为512×384的原始自然内容图像）、*LIVE*（Sheikh，2005）（29幅分辨率各异的原始自然内容图像）、*TID2013*（Ponomarenko等人，2015）（25幅尺寸为 512×384的原始自然内容图像）、*SCID*（Yang、Fang和Lin，2015）（40幅尺寸为 1280×720的原始屏幕内容图像）、*SIQAD*（Ni等人，2017）（20幅分辨率各异的原始屏幕内容图像），以及*MCL-JCI*（Jin等人，2016）（50幅尺寸为 1920×1080的原始自然内容图像）。

评估指标。更好的 JND方法能够隐藏更多噪声。为了评估噪声容忍性能，我们根据预测的可见性阈值，将噪声随机加入每个像素。

$$I_{con} = I_{ori} + \alpha \times \gamma \times I_{vt}, \tag{13}$$

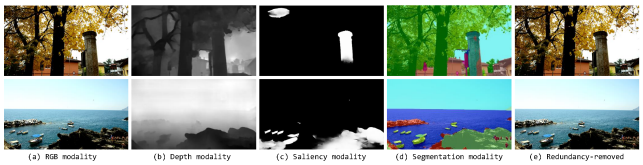


图7：**实验中使用的四种模态示例**。最右侧一列是 I_{rr}的真实值。

其中， I_{con} 表示 JND 污染图像， I_{ori}表示原始图像。 α 表示用于调整噪声水平的调整因子。 γ 表示一个与 I_{ori}大小相同的随机矩阵，其值为 1 或 -1。 I_{vt}表示相应方法预测的可见性阈值。在实验中， I_{vt} 表示可见性阈值，即 I_{ori} 与 I_{rr}之间的差值。

在实验中，我们调整式（13）中的 α ，以确保所有 JND 污染图像具有相同的噪声水平（例如，均方误差（MSE）=100）。为更好地展示比较性能，我们在图像层面评估可见性阈值的预测准确度。具体而言，我们采用峰值信噪比（PSNR）和结构相似性（SSIM）作为基准指标。我们将冗余去除图像与真实图像之间的 PSNR 结果记为 $PSNR_{rr}^{gt}$，SSIM 结果记为 $SSIM_{rr}^{gt}$。

我们采用多尺度结构相似性（MS-SSIM）和平均意见分（MOS）来衡量不同模型生成的污染图像的视觉质量（Zhai 和 Min，2020）。MS-SSIM 是一种客观的全参考图像质量指标，分值范围为 [0,1]。 MS_{con}^{ori} 表示 JND 污染图像与原始图像之间的 MS-SSIM。相应地，MOS 用于评估视觉质量差异，记为 MOS_{con}^{ori}，该分值通过主观实验获得。

总而言之， $PSNR_{rr}^{gt}$和 $SSIM_{rr}^{gt}$衡量去冗余图像的预测准确性。 MS_{con}^{ori}和 MOS_{con}^{ori}衡量 JND 污染图像的感知质量。对于这四项指标，值越高表示结果越好。

主观测试。 主观实验主要依据 ITU-R BT.500-13 标准进行。我们邀请了 20 名受试者参与主观测试，其中男性 12 名、女性 8 名，年龄在 18 至 40 岁之间。两幅图像并排显示在一台 65 英寸显示器上，其中一幅是原始图像，另一幅是 JND 污染图像。主观测试流程与（Shen 等人，2020）中的相同。获得主观结果后，我们采用（Shen 等人，2020）提出的基于统计的异常值剔除方法，检验实验原始结果的一致性。剔除异常值数据后，将平均结果作为下一节中的最终平均意见分（MOS）。

实现细节。 hmJND-Net基于*PyTorch* 实现，所有权重均采用截断正态初始化器进行初始化，并使用默认参数的*Adam* 优化器（例如， β_1 = 0.9和 β_2 = 0.999）。对于HMFA模块，我们执行了3组STA块。我们

表 1：所提出的 hmJND-Net 中三种额外模态的消融研究。

显著性	深度	分割	$PSNR_{rr}^{gt}[+]$	$SSIM_{rr}^{gt}[+]$
✗	✗	✗	33.8645	0.9571
✓	✗	✗	34.3585	0.9634
✗	✓	✗	34.3719	0.9637
✗	✗	✓	34.1246	0.9631
✓	✓	✓	**34.5038**	**0.9652**

表 2：所提出的 HMPF 和 HMFA 模块的消融研究。

HMPF	HMFA	$PSNR_{rr}^{gt}[+]$	$SSIM_{rr}^{gt}[+]$
✗	✗	33.8348	0.9594
✗	✓	34.2866	0.9633
✓	✗	34.0914	0.9608
✓	✓	**34.5038**	**0.9652**

在*Nvidia TeslaA100* GPU上以16为小批量大小训练 hmJND-Net，共训练200轮，耗时约10小时。初始学习率为1e-4，并从第200轮 开始线性衰减至0。由于数据集中的图像尺寸较大（1920×1080），我们对所有输入图像进行自适应划分，划分尺寸为224×224。除上述操作外，我们未使用任何额外的特殊数据增强方法。

消融研究

为验证所引入的先验模态的必要性，以及所提出的多模态融合与对齐模块的有效性，我们在基准数据集上进行了消融研究（Shen 等人，2020）。表 1 和表 2 中，符号“✓”和“✗”分别表示选中和未选中。

三种模态的影响。 为验证所引入的三种模态的必要性，我们进行了五项相关实验，结果列于表 1。在实验中，某一模态的替代方式是重复使用已启用的模态。实验结果表明，同时使用三种模态时效果最佳，这验证了所引入模态的必要性。

HMPF 和 HMFA 的影响。 为验证 HMPF 和 HMFA 模块的有效性，我们还进行了四项补充实验，结果见表 2。在实验中，HMPF 的替代方案是将所有先验模态直接拼接在一起，然后使用一个三层 3×3卷积和一个单层 1×1卷积来减少特征数量。

HMFA 的替代方案是先将先验多模态特征与 RGB 模态拼接起来，然后使用基于 SENet 的通道注意力机制和单层 1×1 卷积对它们进行对齐。实验结果表明，最佳性能可以

在同时启用这两个模块时实现，这验证了所提出的 HMPF 和 HMFA 模块的有效性。

整体性能比较

定性分析。图6展示了 SHEN2020数据集上“I05”的视觉比较，其中图6 (a)为原始图像，图6 (b)-(j)给出了九种方法预测的视觉冗余结果，其中较亮的像素表示更高的冗余。各方法（例如，*Yang2005*）的详细信息见表3。

如图所示，现有的代表性方法可能会在显著物体或平坦区域预测出较高的可见性阈值，从而导致视觉质量较差。主要原因是平滑和均匀区域更为敏感，细微变化也可能引起人类视觉系统的注意。得益于多模态信息的联合引导，hmJND-Net 会在纹理区域（例如树木和草地）容忍更多噪声，因为这些区域不易被人类视觉系统察觉；而在平滑、均匀且敏感的区域则容忍较少噪声。因此，如图6 (j)所示，我们的 hmJND-Net 能更好地预测冗余。

定量分析。如第节所述，我们在八个广泛使用的图像数据集上开展了定量实验。实验中，根据式（13），所有测试图像的噪声注入水平均为 MSE=100 。

表3给出了九种代表性模型在八个基准数据集上的 MS^{ori}_{con} 结果。可以看出，在所有八个数据集上，在相同噪声水平下，所提出的方法均取得了最高的平均多尺度结构相似性结果。

此外，我们还在 SHEN2020数据集上，以 MOS^{ori}_{con} 为指标，在相同噪声水平（MSE=100）下开展了主观实验。各模型（表3中从上到下）的平均 MOS^{ori}_{con} 结果依次为 -1.5071、-1.7214、-1.5929、-1.1429、-1.0572、-1.8964、-1.1842、-1.0833 和 -0.5536。由此可见，我们的 hmJND-Net 获得了最高的平均意见分。

JND引导的压缩性能

本节将我们的 hmJND-Net 集成到压缩应用中，包括广泛使用的 JPEG，以及两种最先进的压缩标准（例如 HEVC 和 VVC）。

JND引导的JPEG压缩。在 JND引导的JPEG压缩中，一种广泛使用的方法是在编码前去除输入图像中的视觉冗余，从而在保持相同视觉质量的同时节省比特率。具体而言，每个 8×8块都按如下方式通过预测的可见性阈值进行处理:

$$\widehat{I}_{ori}(\mathbf{p}) = \begin{cases} \overline{I}_b, & if\ |I_{ori}(\mathbf{p}) - \overline{I}_b| \le I_{vt}(\mathbf{p}), \\ I_{ori}(\mathbf{p}) + I_{vt}(\mathbf{p}), & if\ I_{ori}(\mathbf{p}) - \overline{I}_b < -I_{vt}(\mathbf{p}), \\ I_{ori}(\mathbf{p}) - I_{vt}(\mathbf{p}), & otherwise. \end{cases}, \quad (14)$$

其中， $I_{ori}(\mathbf{p})$ 表示位置$\mathbf{p}$处的像素幅值，$I_{vt}(\mathbf{p})$ 表示相关的视觉阈值，I_b 表示当前 8×8块的平均亮度值，$\widehat{I}_{ori}(p)$表示预处理后的输出。



图 8：**三种不同编码平台上四种先进 JND 引导方法的视觉效果对比**。

JND 引导的 HEVC/VVC 压缩。在 HEVC 或 VVC 标准中，我们首先获得 JND 引导残差值，然后由 HEVC 或 VVC 编码器对残差块进行编码。具体而言，我们按如下方式计算 JND 引导残差值:

$$\widehat{R}(\mathbf{p}) = \begin{cases} 0, & if\ |R(\mathbf{p})| \le I_{vt}(\mathbf{p})\ and\ \sigma^2(\mathbf{p}) > \sigma^2, \\ \min(R(\mathbf{p}) \times \frac{|I_{vt}(\mathbf{p})|}{\sigma^2},\ R(\mathbf{p}) + I_{vt}(\mathbf{p})), & else\ if\ R(\mathbf{p}) < 0, \\ \max(R(\mathbf{p}) \times \frac{|I_{vt}(\mathbf{p})|}{\sigma^2},\ R(\mathbf{p}) - I_{vt}(\mathbf{p})), & otherwise. \end{cases}, \quad (15)$$

其中， $\sigma^2(\mathbf{p})$ 和 σ^2 分别表示位置 $\mathbf{p}$ 处局部范围和编码块的相关方差， $R(\mathbf{p})$ 表示通过 HEVC 或 VVC 获得的残差值，而 $\widehat{R}(\mathbf{p})$ 表示处理后的预测残差值。

压缩性能比较。我们在 *SHEN2020* 上进行了压缩实验，并根据比特/像素（bpp）对四种代表性方法进行了比较。例如，JPEG 无损编码、(Chen and Wu 2019) (*Chen19JPEG*)、(Wang et al. 2022) (*Wang22JPEG*)、(Jiang et al. 2022) (*Jiang22JPEG*) 和我们的 hmJND-Net (*OursJPEG*) 的平均 bpp 结果分别为 4.3031、3.4440、3.5604、4.0643 和 3.3771。HEVC 编码器上的 HEVC 无损编码、*Chen19HEVC*、*Wang22HEVC*、*Jiang22HEVC* 和 *Our-sHEVC* 的平均 bpp 结果分别为 5.0411、3.6557、3.6366、4.2253 和 3.4144。VVC 编码器上的 VVC 无损编码、*Chen19VVC*、*Wang22VVC*、*Jiang22VVC* 和 *OursVVC* 的平均 bpp 结果分别为 4.5304、2.9958、2.7764、3.0934 和 2.6328。在视觉质量相近的情况下，我们的 hmJND 引导编码方法平均可节省 32.3809% 的比特率。

此外，图 8 展示了四种方法在“I05”上的视觉压缩性能对比。可以看出，我们采用 hmJND-Net 引导的压缩结果与无损编码方案的视觉质量几乎相同，而我们采用 hmJND 引导的方法平均可节省 28.4016% 的比特率。

结论

本文提出了一种基于同源多模态

表3：8个基准数据集上9种代表性方法的平均 MS_{con}^{ori} 结果。每行中的最佳结果以**粗体**显示。

方法 \ 数据集	*CSIQ*	*KADID-10K*	*LIVE*	*TID2013*	*SCID*	*SIQAD*	*MCL-JCI*	*SHEN2020*	**平均值**
Yang2005 (Yang 等，2005)	0.9591	0.9513	0.9543	0.9519	0.9525	0.9686	0.9352	0.9510	0.9530
Liu2010 (Liu 等，2010)	0.9629	0.9547	0.9576	0.9548	0.9558	0.9712	0.9370	0.9495	0.9554
Wu2013(Wu 等，2013)	0.9598	0.9541	0.9551	0.9520	0.9676	0.9831	0.9365	0.9504	0.9573
Wu2017 (Wu 等，2017)	0.9647	0.9617	0.9629	0.9599	0.9650	0.9761	0.9490	0.9593	0.9623
Chen2019 (Chen and Wu 2019)	0.9690	0.9659	0.9658	0.9630	0.9691	0.9765	0.9567	0.9634	0.9662
Shen2020(Shen 等人，2020)	0.9558	0.9454	0.9477	0.9452	0.9463	0.9609	0.9301	0.9418	0.9466
Wang2022(Wang 等人，2022)	0.9677	0.9613	0.9616	0.9586	0.9677	0.9820	0.9412	0.9581	0.9623
Jiang2022 (Jiang et al. 2022)	0.9599	0.9689	0.9620	0.9591	0.9724	0.9772	0.9576	0.9627	0.9650
Ours	**0.9709**	**0.9737**	**0.9695**	**0.9666**	**0.9795**	**0.9880**	**0.9584**	**0.9710**	**0.9722**

学习的端到端视觉冗余预测网络。具体而言，我们考虑了三种先验模态，从不同角度丰富训练数据。为有效利用多模态信息，我们开发了同源模态先验融合（HMPF）模块，基于求和增强和减法偏移融合先验模态特征。此外，我们提出了同源模态特征对齐（HMFA）模块，基于自注意力驱动的编码器-解码器模块对齐来自不同领域的中间特征。在八个不同基准数据集上的实验结果从定性和定量两方面证明了该方法的优越性。我们相信，所提出的多模态驱动方法将有助于提升图像压缩和处理系统的服务质量。

参考文献

Bae, S.-H.; Kim, J.; and Kim, M. 2016. 基于 HEVC 的感知自适应视频编码：采用基于 DCT 的局部失真检测概率模型。*IEEE*图像处理汇刊, 25(7): 3343–3357.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. BEiT：图像Transformer的BERT预训练。载于国际学习表征会议（*ICLR*），1469–1479。

Chen, Z.; and Wu, W. 2019. 面向具有视野不均匀性的图像的非对称中央凹恰可察觉差异模型。*IEEE*视频技术电路与系统汇刊, 30(11): 4064–4074. Cheng, X.; Fu, Z.; and Yang, J. 2020. 基于深度引导的内部退化学习的零样本图像超分辨率。发表于 施普林格欧洲计算机视觉会议（*ECCV*）, 265–280. Chou, C.-H.; and Li, Y.-C. 1995. 基于恰可察觉失真轮廓度量的感知调优子带图像编码器。*IEEE*视频技术电路与系统汇刊, 5(6): 467–476. Hu, J.; Shen, L.; and Sun, G. 2018. 挤压激励网络。发表于*IEEE*计算机视觉与模式识别会议（*CVPR*）, 7132–7141. Huang, Q.; Wang, H.; Lim, S. C.; Kim, H. Y.; Jeong, S. Y.; and Kuo, C.-C. J. 2017. HEVC感知有损/无损边界QP值的度量与预测。发表于*IEEE*数据压缩会议（*DCC*）, 42–51.

Jiang, Q.; Liu, Z.; Wang, S.; Shao, F.; and Lin, W. 2022. 面向自然图像的自顶向下恰可察觉差异估计。*IEEE*图像处理汇刊, 1608–1612.

Jin, L.; Lin, J. Y.; Hu, S.; Wang, H.; Wang, P.; Katsavouni- dis, I.; Aaron, A.; and Kuo, C.-C. J. 2016. 基于MCL-JCI数据集的构建与分析，对感知JPEG图像质量进行统计研究。*ElectronicImaging*, 2016(13)：1–9。

Ki, S.; Bae, S.-H.; Kim, M.; and Ko, H. 2018. 用于感知视频编码的基于学习的可觉察量化失真建模。*IEEE Transactionson Image Pro-cessing*，27(7)：3178–3193。

Larson, E. C.; and Chandler, D. M. 2010. 最显著失真：全参考图像质量评估与策略的作用。*SPIE* 电子成像期刊, 19(1): 1116–1137.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k：一个大规模人工失真图像质量评估数据库。载于国际多媒体体验质量会议 *(QoMEX)*, 1–3.

Liu, A.; Lin, W.; Paul, M.; Deng, C.; and Zhang, F. 2010. 基于分解模型分离边缘和纹理区域的图像恰可察觉差异。*IEEE* 电路与系统视频技术汇刊, 20(11): 1648–1652.

Liu, H.; Zhang, Y.; Zhang, H.; Fan, C.; Kwong, S.; Kuo, C.-C. J.; and Fan, X. 2019. 一种基于深度学习的逐图像恰可察觉失真预测模型，用于图像压缩。*IEEE* 图像处理汇刊, 29: 641–656.

Liu, X.; Zhan, X.; and Wang, M. 2020. 一种新颖的基于边缘模式的屏幕内容图像恰可察觉差异模型。发表于 *IEEE InternationalConference on Signaland Image Processing (ICSIP)*，386–390。

Liu, Y.; Liu, J.; Argyriou, A.; Ma, S.; Wang, L.; and Xu, Z. 2021a. 基于球面小波变换的360度虚拟现实视频水印。*ACM TransactionsonMultimedia Computing, Communications,and Applications*，17(1)：1–23。

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer：分层视觉

采用移位窗口的 Transformer。在*IEEE*国际计算机视觉会议 *(ICCV)*，10012–10022。Min, X.; Zhai, G.; Zhou, J.; Farias, M. C.; and Bovik, A. C. 2020。视听信号主观与客观质量评估研究。*IEEE*图像处理汇刊，29：6054–6068。Ni, Z.; Ma, L.; Zeng, H.; Chen, J.; Cai, C.; and Ma, K.-K. 2017。ESIM：用于屏幕内容图像质量评估的边缘相似度。*IEEE*图像处理汇刊，26(10)：4818–4831。Peterson, H. A.; and Ahumada Jr, A. J. 1992。基于亮度模型的彩色图像压缩 DCT 量化。载于人类视觉、视觉处理与数字显示 *III*，第1666卷，365–374。Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazar- ian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; 等。2015。图像数据库 TID2013：特点、结果与展望。*Elsevier*《信号处理：图像通信》，30：57–77。Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019。BASNet：边界感知显著目标检测。载于*IEEE*计算机视觉与模式识别会议 *(CVPR)*，7479–7489。Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021。用于密集预测的视觉 Transformer。载于*IEEE*国际计算机视觉会议 *(ICCV)*，12179–12188。Seo, S.; Ki, S.; and Kim, M. 2020。一种新颖的基于恰可察觉差异的显著性通道注意力残差网络，用于全参考图像质量预测。*IEEE*视频技术电路与系统汇刊，31(7)：2602–2616。Shao, H.; and Wang, Y. 2021。结合显著性先验与相对总变分的生成式图像修复。爱思唯尔《视觉通信与图像表示》，79：101–110。Sheikh, H. 2005。LIVE图像质量评估数据库第2版。*http://live.ece.utexas.edu/research/quality*。Shen, X.; Ni, Z.; Yang, W.; Zhang, X.; Wang, S.; and Kwong, S. 2020。恰可察觉失真轮廓推断：一种块级结构可见性学习方法。*IEEE*图像处理汇刊，30：26–38。Tian, T.; Wang, H.; Kwong, S.; and Kuo, C.-C. J. 2021。基于块级恰可察觉差异预测的感知图像压缩。*ACM*多媒体计算、通信与应用汇刊，16(4)：1–15。Wang, H.; Gan, W.; Hu, S.; Lin, J. Y.; Jin, L.; Song, L.; Wang, P.; Katsavounidis, I.; Aaron, A.; and Kuo, C.-C. J. 2016。MCL-JCV：基于 JND 的 H.264/AVC 视频质量评估数据集。载于*IEEE*国际图像处理会议 *(ICIP)*，1509–1513。Wang, H.; Wang, L.; Hu, X.; Tu, Q.; and Men, A. 2014。基于显著性与恰可察觉失真的 H.265/HEVC 感知视频编码。载于无线个人多媒体通信国际研讨会 *(WPMC)*，106–111。Wang, M.; Xu, Z.; Liu, X.; Xiong, J.; and Xie, W. 2022. 基于可见性建模和深度预测的屏幕内容数据感知准无损压缩。*IEEE*工业信息学汇刊, 6865–6875. Wang, X.; Yu, K.; Dong, C.; and Loy, C. C. 2018. 通过深度空间特征变换恢复图像超分辨率中的逼真纹理。载于 *IEEE*计算机视觉与模式识别会议 *(CVPR)*, 606–615. Wei, Z.; and Ngan, K. N. 2009. DCT 域中灰度图像/视频的时空恰可察觉失真轮廓。*IEEE*视频技术电路与系统汇刊, 19(3): 337–346. Wu, J.; Li, L.; Dong, W.; Shi, G.; Lin, W.; and Kuo, C.-C. J. 2017. 针对图案复杂度较高图像的增强型恰可察觉差异模型。*IEEE*图像处理汇刊, 26(6): 2682–2693. Wu, J.; Shi, G.; Lin, W.; Liu, A.; and Qi, F. 2013. 基于自由能原理的图像恰可察觉差异估计。*IEEE*多媒体汇刊, 15(7): 1705–1710. Xie, W.; Huang, T.; and Wang, M. 2022. MNSRNet：用于三维表面超分辨率的多模态变换器网络。载于*IEEE*计算机视觉与模式识别会议 *(CVPR)*, 12703–12712. Yang, H.; Fang, Y.; and Lin, W. 2015. 屏幕内容图像的感知质量评估。*IEEE*图像处理汇刊, 24(11): 4408–4421. Yang, X. K.; Lin, W. S.; Lu, Z. K.; Ong, E. P.; and Yao, S. S. 2005. 恰可察觉失真模型及其在视频编码中的应用。爱思唯尔《信号处理：图像通信》, 20(7): 662–680. Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. Ernie-vil：通过场景图增强知识的视觉语言表征。载于*AAAI*人工智能会议, 第35卷, 3208–3216. Zhai, G.; and Min, X. 2020. 感知图像质量评估：综述。施普林格《中国科学：信息科学》, 63(11): 1–52. Zhang, X.; Wang, S.; Gu, K.; Lin, W.; Ma, S.; and Gao, W. 2016. 基于恰可察觉差异的 JPEG 压缩感知优化。*IEEE*信号处理快报, 24(1): 96–100. Zhang, X. H.; Lin, W. S.; and Xue, P. 2005. 恰可察觉视觉失真的改进估计。爱思唯尔《信号处理》, 85(4): 795–808.