全景分割与多视觉特征协同的图像描述生成方法

刘明明[1,2]，陆劲夫[2]，刘浩[2]，张海燕[1]

（1. 江苏建筑职业技术学院智能制造学院，江苏 徐州 221116；2. 中国矿业大学计算机科学与技术学院，江苏 徐州 221116）

摘　要： 现有基于 Transformer 架构的图像描述生成模型取得了较好的泛化性能，然而，大多数方法通常使用区域视觉特征进行编解码，导致无法全面利用整幅图像的细粒度信息，且存在视觉特征混淆问题。为此，将全景分割引入图像描述生成过程，使用基于全景分割的掩膜视觉特征代替区域视觉特征，提出一种全景分割与多视觉特征协同的图像描述生成方法。该方法不仅可以有效解耦视觉表征，而且能够充分结合掩膜视觉特征和网格视觉特征的优势，提升图像描述生成的可解释性和描述性能。在 MSCOCO 标准数据集上进行定量和定性实验，结果表明，所提方法不仅可以显著提升现有模型的性能，同时能够增强图像描述生成过程的可解释性，CIDEr 和 BLEU-4 指标分别达到 138.5 和 41。

关键词： 图像理解；图像描述生成；全景分割；特征融合；视觉编码

中图分类号： TP391　　**文献标志码：** A　　**DOI：** 10.19678/j.issn.1000-3428.0069303

Image Description Generation Method by Panoptic Segmentation and Multi-Visual-Feature Fusion

LIU Mingming[1,2], LU Jinfu[2], LIU Hao[2], ZHANG Haiyan[1]

(1. School of Intelligent Manufacturing, Jiangsu Vocational Institute of Architectural Technology, Xuzhou 221116, Jiangsu, China;
2. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, Jiangsu, China)

【Abstract】 Due to their powerful sequence modeling capabilities, Transformer-based image captioning models have demonstrated remarkable performance. However, most of these models typically utilize region visual features to perform encoding and decoding, which cannot fully use the fine-grained information of the whole image, and this leads to visual feature confusion. Accordingly, we introduce panoptic segmentation into the Transformer-based image captioning model by replacing the region visual feature with mask visual features and propose a novel image captioning model based on multi-visual-feature fusion. Our model not only disentangles the region visual features effectively but also makes use of both mask and grid visual features to improve image captioning performance. We perform quantitative and qualitative experiments on the MSCOCO dataset, which demonstrate that our method significantly outperforms existing Transformer-based image captioning models. In addition, our model enhances the interpretability of the caption generation process, and more specifically, achieves CIDEr and BLEU-4 scores of 138.5 and 41, respectively.

【Key words】 image understanding; image description generation; panoptic segmentation; feature fusion; visual encoding

0　引言

图像描述生成作为图像理解研究领域的基础性研究课题[1-3]，旨在让机器理解图像的内容，并且以人类语言的形式自动生成对应的描述语句。在图像检索和分类领域，图像描述用于提升图像内容检索和分类的准确度[4]；在视觉辅助领域，图像描述帮助视力障碍人群理解图像信息，弥补视觉缺陷[5]；在智慧医疗领域，图像描述可以自动生成医学图像诊断报告，为智能诊疗提供技术支撑[6]。

近年来，基于 Transformer 的编码器-解码器架构已经成为图像描述的主流模型，这类模型通常使用目标检测模型（例如 Faster R-CNN）来提取图像的区域级视觉特征，然后将其输入编码器-解码器网络来生成相应的文本描述。然而，基于目标检测模型提取的区域视觉特征存在以下问题：区域特征通常聚焦图像中的前景目标，无法全面表征整幅图像的细粒度信息；多个区域视觉特征之间存在特征混淆现象。为了解决这些问题，近期研究人员先后引入了语义概念、网格特征、深度估计等其他视觉线

收稿日期： 2024-01-26　　**修回日期：** 2024-04-03

基金项目： 国家自然科学基金（61801198）；江苏省自然科学基金（BK20180174）。

通信作者 E-mail： jsjzi_lmm@126.com

索，通过与 Faster R-CNN 提取的区域特征相结合来缓解上述问题。然而，这些方法仍未从根本上解决传统区域视觉特征的局限性，从而制约了图像描述性能的进一步提升。

本文首先将全景分割引入到图像视觉特征表示中，用基于全景分割的掩膜视觉特征来代替区域视觉特征。此外，可以通过全景分割提取图像中的背景信息，提供更全面的图像全局信息。在此基础上，将掩膜视觉特征和网格视觉特征进行融合，提出一种新的多视觉特征融合的图像描述生成方法。该方法通过有效解耦视觉表征以及充分利用掩膜视觉特征和网格视觉特征的优势，提升图像描述生成的可解释性和描述性能。在解码阶段，特征融合模块使每个特征可以独立学习其在文本预测时的权重，更有效地利用视觉信息流中的丰富线索，增强模型在生成描述时的准确性和全面性。在 MSCOCO 标准数据集上进行定量和定性实验，以验证所提方法的有效性。本文的主要贡献包括以下 2 个方面：

1）设计了一种基于全景分割的掩膜视觉特征提取方法，不仅可以精确表征图像前景目标视觉特征，而且能够有效表征整幅图像的细粒度信息。

2）提出一种新的基于 Transformer 的多视觉特征融合图像描述生成模型，利用掩膜视觉特征和网格视觉特征的互补性，提升视觉和文本特征之间的多模态交互能力。

1 相关工作

近年来，自然语言处理中的 Transformer 架构开始被引入图像描述领域，并取得了比 CNN、RNN 模型更好的图像描述性能。基于 Transformer 框架，国内外研究者进行了大量的改进工作。文献[7]提出了几何注意力机制，使得模型能够在对图像编码的过程中关注到目标对象在空间上的几何信息。与之工作相近的是，文献[8]提出了一种规范化的几何感知自注意力模型，其能够更加有效地考虑图像中目标对象之间的几何关系。文献[9]使用 2 个独立的 Transformer 编码器分别对编码视觉特征和语义属性信息进行编码，在解码端引入纠缠注意力以弥补传统注意力在两者之间缺乏的互补性，并设计门控机制对视觉信息和语义信息进行选通，使其具备同时关注视觉特征和语义属性信息的能力。文献[10]将一组记忆向量与 Transformer 编码器进行集成，在丰富图像区域之间多层次视觉关系的同时学习和编码先验知识，并利用网状结构以完全连接的方式将编码器连接到解码器，使得解码器可以充分利用低高层次的视觉关系。

由于传统的注意力机制往往利用线性融合的方式学习跨模态的特征交互，其本质只挖掘了模态间一阶的特征交互，极大限制了注意力机制在图像描述生成等跨模态内容推理任务中的应用。因此，文献[11]提出 X 线性注意力，通过利用双线性池化技术（Bilinear Pooling）来捕捉两阶特征之间的相互作用，以增强输出特征的表征能力。文献[12]设计了一个全局增强型编码器来学习全局特征，并设计了一个全局自适应解码器进行全局特征的嵌入。文献[13]提出了一种融合视觉区域和网格特征的方法，使用交叉注意力模块将 2 种特征进行交互融合。

近年来，国内学者也开展了相关研究。文献[14]提出结合视觉特征和场景语义的图像描述生成方法，利用潜在狄利克雷分布模型与多层感知机提取图像场景语义相关的主题词，通过主题词指导单词的准确生成。文献[15]提出基于强化学习的多层级视觉融合网络模型，通过将视觉特征转化为视觉知识的特征集，从而生成更加流畅的描述语句。文献[16]利用视觉关联与上下文双注意力机制，指导生成准确的图像描述文本。文献[17]通过视觉区域聚合与双向协作学习，促进模型生成更加细粒度的描述文本。

尽管上述模型获得了比传统方法更优的描述准确性，但这些模型大多利用区域视觉特征进行编解码，受限于视觉表征的局限性，模型的可解释性和泛化性能还有待提升。

传统的深度语义分割和实例分割任务单独设计网络架构，分别进行逐像素的分类和基于掩码的分类。近年来，以 MaskFormer[18] 为代表的模型利用掩码分类统一了语义分割和实例分割任务，同时提供全景分割能力。之后的工作也大多针对提取的掩码特征进行改进，例如 Mask2Former[19] 提出一种掩码注意力机制，使每个对象查询只关注预测掩码的前景区域，kMaX-DeepLab[20] 从交叉注意力学习重新表述为聚类的过程，将对象查询作为具有可学习嵌入向量的聚类中心来预测最终的掩码嵌入。这些方法通过对掩码特征进行语义表征，精确地提取了图像中的语义信息，从而实现分割领域的统一和性能提升。和本文工作相似，文献[21]也利用了全景分割特征来补充网格特征，但其目的是提供每个网格的语义类别来作为空间语义指导。与之不同，本文将全景分割提取的掩膜特征作为区域特征来与网格特征进行融合，利用全景分割提取的掩码特征为

图像理解提供更丰富的信息，从而通过有效解耦视觉表征以及充分利用掩膜视觉特征和网格视觉特征的优势，提升图像描述生成的可解释性和描述性能。

2 本文方法设计与实现

本文提出基于全景分割与多视觉特征融合的 Transformer 图像描述生成模型，如图 1 所示，该模型首先利用网格特征提取器和掩膜特征提取器，分别提取网格特征和基于全景分割的掩膜特征。由于网格和掩膜视觉特征具有不同的视觉表征特性，因此 2 种特征使用不共享参数的 Transformer 编码器分别进行编码。然后，引入融合模块将掩膜特征和网格特征相结合，为解码端提供丰富的视觉信息。

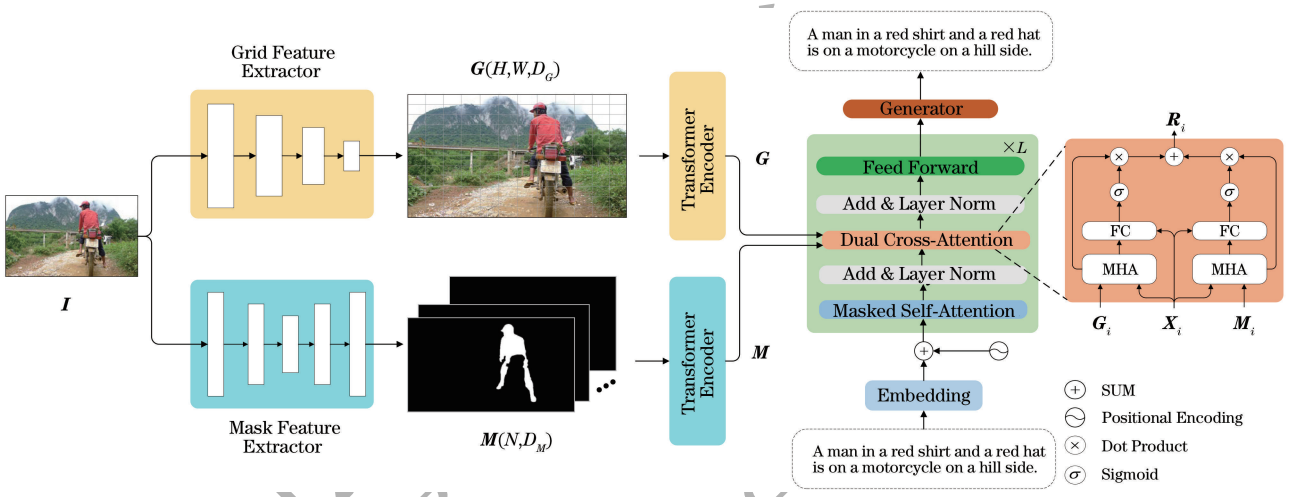


图 1 模型总体架构

Fig. 1 Overall architecture of the model

在本节中，首先介绍网格特征提取过程，然后提出一种基于全景分割的掩膜特征提取方法，接着设计对 2 种视觉特征进行特征融合的模块，最后介绍所提模型的训练和推断过程。

2.1 Transformer 编解码模块

本文方法仍然基于流行的 Transformer 模型进行构建。Transformer 架构完全通过注意力机制对序列依赖关系进行建模，以解决长序列依赖问题，使得模型可以并行化计算，大大提高了预测精确度和训练速度。

如图 1 所示，Transformer 模型遵循编码-解码架构，左侧为网格特征与掩膜特征编码器，包含 L 个编码器块，右侧为解码器，同时由 L 个解码器块组成，两者都包含了多头注意力(MHA)、残差和归一化(Add & Norm)、前馈网络(FFN)等重要组成部分。此外，解码部分引入位置编码(PE)操作，这是因为输入文本是按一定顺序来排列的，通过位置编码可以有效地学习到单词之间的相对位置。而在编码部分，因为图像的不同区域之间并没有固定的顺序关系，图像特征编码序列通过 Transformer 全注意力机制建模特征之间的关系，而无须引入位置编码。

如图 2 所示，多头注意力将高维空间的查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 以及值向量矩阵 $\boldsymbol{V}$ 划分为 h 个不同的子空间分别计算相似度，保证每个子空间的独立性，最后采用变换方法进行合并，计算公式如下：

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{W}_O[\boldsymbol{h}_{\mathrm{head},1},\cdots,\boldsymbol{h}_{\mathrm{head},h}] \tag{1}$$

$$\boldsymbol{h}_{\mathrm{head},i}=\mathrm{Attention}(\boldsymbol{U}_i\boldsymbol{Q},\boldsymbol{X}_i\boldsymbol{K},\boldsymbol{Y}_i\boldsymbol{V}) \tag{2}$$

式中：$\boldsymbol{U}_i,\boldsymbol{X}_i,\boldsymbol{Y}_i\in\mathbb{R}^{d\times d}$ 是可学习的参数，将各个矩阵映射到第 i 个 $d_k(d_k=d/h)$ 维的子空间中；$[\boldsymbol{h}_{\mathrm{head},1},\cdots,\boldsymbol{h}_{\mathrm{head},h}]$ 代表将 h 个不同子空间进行拼接，随后通过可学习的线性变换 $\boldsymbol{W}_O\in\mathbb{R}^{d\times d}$ 得到最终的输出。多头注意力有利于从不同子空间中获取更丰富的信息，进而增强模型的表征能力。

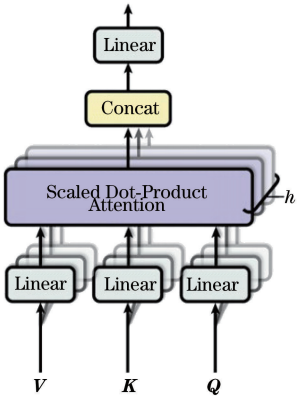


图 2 多头注意力

Fig. 2 Multi-head attention

此外，根据查询矩阵 $\boldsymbol{Q}$ 、键矩阵 $\boldsymbol{K}$ 以及值矩阵 $\boldsymbol{V}$ 是否源自同一序列的情况，多头注意力又可分为自注意力（Self-Attention）和交叉注意力（Cross-Attention）。如图 1 的 Transformer 模型所示，编码器和解码器的第一个多头注意力输入的矩阵 $\boldsymbol{Q}$ 、$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 都来自同一个特征向量，因此可以称之为自注意力。而解码器部分第二个多头注意力的查询矩阵 $\boldsymbol{Q}$ 来自于编码器的最终输出，与键矩阵 $\boldsymbol{K}$ 以及值矩阵 $\boldsymbol{V}$ 的来源并不相同，因此称之为交叉注意力。一般情况下，交叉注意力常用于不同模态数据之间的交互，对于图像描述生成这类跨模态任务起着相当重要的作用。

Transformer 模型的另外一个组成部分是前馈网络 FFN，它的主要作用是提供非线性变换，提高模型对复杂过程的拟合能力，计算公式如式（3）所示，可以看出前馈网络将多头注意力的输出特征作为输入，并通过 2 个全连接层（FC）对特征向量进行非线性变换，此外，ReLU 激活函数和随机失活（Dropout）介于两层之间。

$$\begin{aligned}&\mathrm{FFN}(\boldsymbol{x})=\\&\mathrm{FC}(\mathrm{Dropout}(\mathrm{ReLU}(\mathrm{FC}(\boldsymbol{x},4d),t),d))\end{aligned}\tag{3}$$

式中：输入特征 $\boldsymbol{x}$ 的维度为 d，它首先被投影到 $4d$ 维，然后被映射到 d 维；t 是随机失活率。

每个子层后都会额外增加一个残差与归一化层（Add & Norm），它主要应用于多头注意力层和前馈网络层之后。与循环神经网络和卷积神经网络等顺序算法不同，Transformer 模型中的多头注意力虽然关注到了单词之间的重要程度，但无法捕捉到单词间相对或绝对的位置信息。为解决这一问题，Transformer 模型额外将位置编码 PE 添加到输入的特征向量中。位置编码采用正弦和余弦函数获取单词的顺序位置信息，其结果随后与词嵌入向量相加，作为输入送至多头注意力。位置编码的定义如下：

$$\begin{aligned}&\mathrm{PE}(d_{\mathrm{pos}},2i)=\sin\left(\frac{d_{\mathrm{pos}}}{10\ 000^{\frac{2i}{d}}}\right)\\&\mathrm{PE}(d_{\mathrm{pos}},2i+1)=\cos\left(\frac{d_{\mathrm{pos}}}{10\ 000^{\frac{2i}{d}}}\right)\end{aligned}\tag{4}$$

式中：d_{pos} 代表当前单词在序列中的位置。

2.2 网格视觉特征提取

近年来，一些基于 Transformer 的变体模型尤其是 Swin-Transformer[22] 取得了良好的网格视觉特征提取性能，并成功应用于图像描述领域。为了获得高质量的网格视觉特征并与目前主流方法进行公平对比，本文仍然采用 Swin-Transformer 作为特征提取器。首先使用一个分块模块将 384×384×3 的输入图像分割成不重叠的块，然后每个块被线性嵌入到一个低维空间中生成块嵌入向量，进行四阶段的下采样（分辨率分别为×4、×8、×16、×32），除了第一个阶段使用线性嵌入层外，其他 3 个下采样阶段输出均送入多层 Swin-Transformer 块，这些块使用基于窗口的多头自注意力和位移多头自注意力，最终得到网格特征 $\boldsymbol{G}$，表示如下：

$$\boldsymbol{G}=\mathrm{Gridnet}(\boldsymbol{I}),\ \boldsymbol{G}\in\mathbb{R}^{H\times W\times D_G}\tag{5}$$

式中：Gridnet 表示 Swin-Transformer 网格特征提取网络；H 和 W 分别表示特征图的高度和宽度；D_G 表示特征维度。

2.3 掩膜视觉特征提取网络

与先前的图像描述方法不同，本文所提出的掩膜视觉特征提取网络结合全景分割生成的掩膜表示生成区域视觉特征。掩膜特征提取网络利用全景分割掩膜提取关键的语义特征，不仅可以缓解视觉混淆问题，而且弥补了网格特征表征能力的不足。如图 3 所示，本文使用 kMaX-DeepLab[20] 全景分割网络生成特征图 $\boldsymbol{F}$ 以及掩膜 $\boldsymbol{S}$，该网络从 k-means 聚类的角度重新设计了交叉注意力模块，取得了较好的全景分割性能，同时提供更加精细的分割掩膜。然后，通过在特征图和掩膜上进行池化操作来得到掩膜区域特征 $\boldsymbol{M}$。

具体而言，kMaX-DeepLab 包括像素编码器、像素解码器和 kMaX 解码器 3 个组件。像素编码器作为网络的骨干，负责提取图像特征 $\boldsymbol{F}$。像素解码器包括 Transformer 模块以及用于生成更高分辨率特征的上采样层。图像特征 $\boldsymbol{F}$ 输入到像素解码器以提取其像素特征，同时，一系列 kMaX 解码器将聚类中心转换为掩膜嵌入向量。将 kMaX 解码器学习的掩模嵌入向量与像素特征相乘以产生分割特征 $\boldsymbol{S}$。对图像特征 $\boldsymbol{F}$ 与分割特征 $\boldsymbol{S}$ 进行池化操作生成最终的掩膜特征 $\boldsymbol{M}$。每一个分割区域的视觉特征计算公式如下：

$$\boldsymbol{M}(N,D_M)=\sum_{h,w}\boldsymbol{S}(N,h,w)\cdot\boldsymbol{F}(h,w,D_M)\tag{6}$$

式中：$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N)$，$\boldsymbol{s}_i\in\{0,1\}^{H\times W}$ 表示掩膜矩阵，其中每个元素表示对应的像素点是否属于某一实例对象；N 表示图像中分割区域的个数；D_M 表示每个分割区域对应的视觉特征维度；h 和 w 分别表示每个分割区域对应的行号和列号。

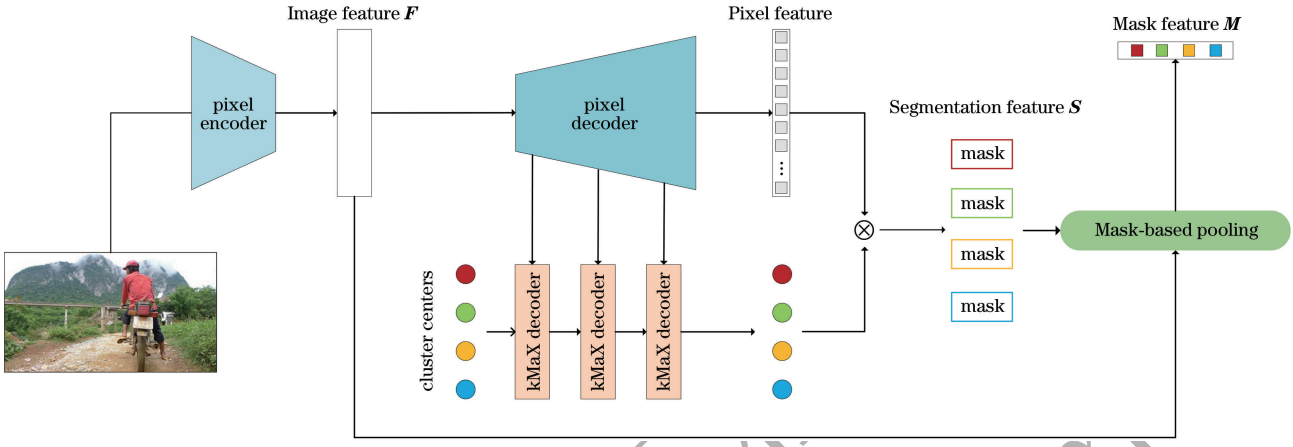


图 3 掩膜视觉特征提取示意图

Fig. 3 Schematic diagram of mask visual feature extraction

2.4 解码网络

如图 1 所示，解码网络采用双流架构，包含已生成描述文本、网格视觉特征和掩膜视觉特征 3 个输入。通过将词嵌入向量和位置编码相加得到单词特征 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$，然后将其送入解码器。在每个解码层中，首先通过自注意力模块对单词特征进行编码，得到相应的语义特征，表示如下：

$$\boldsymbol{X}_i=\text{MultiHead}(\boldsymbol{X}_i,\boldsymbol{X}_i,\boldsymbol{X}_i) \tag{7}$$

$$\boldsymbol{X}_i=\boldsymbol{X}_i+\text{LayerNorm}(\boldsymbol{X}_i) \tag{8}$$

式中：i 表示层索引号；MultiHead(·) 表示 Transformer 多头注意力操作。

然后，通过双流交叉注意力将文本语义特征分别与掩膜特征 $\boldsymbol{M}$ 和网格特征 $\boldsymbol{G}$ 进行交叉融合。将网格视觉特征 $\boldsymbol{G}_i$ 与掩膜视觉特征 $\boldsymbol{M}_i$ 作为键-值分别输入 2 个多头注意力模块中，文本语义特征 $\boldsymbol{X}_i$ 作为查询向量。双流交叉注意力模块分别输出融合特征 $\boldsymbol{g}_i$ 和 $\boldsymbol{m}_i$，其中，$\boldsymbol{m}_i$ 是掩膜视觉特征 $\boldsymbol{M}_i$ 与文本语义特征 $\boldsymbol{X}_i$ 的融合，$\boldsymbol{g}_i$ 是网格视觉特征 $\boldsymbol{G}_i$ 与文本语义特征 $\boldsymbol{X}_i$ 的融合，这种双流交叉学习方式让模型学习到不同特征之间的关联性，提高了特征表达能力和语义理解能力。在此基础上，利用 Sigmoid 激活函数分别计算双流权重 $\boldsymbol{\alpha}_m$ 和 $\boldsymbol{\alpha}_g$，并通过权重对 $\boldsymbol{g}_i$ 和 $\boldsymbol{m}_i$ 进行融合，形成双流融合特征 $\boldsymbol{R}_i$，表示如下：

$$\boldsymbol{g}_i=\text{MultiHead}_1(\boldsymbol{X}_i,\boldsymbol{G}_i,\boldsymbol{G}_i) \tag{9}$$

$$\boldsymbol{m}_i=\text{MultiHead}_2(\boldsymbol{X}_i,\boldsymbol{M}_i,\boldsymbol{M}_i) \tag{10}$$

$$\boldsymbol{\alpha}_m=\sigma(\boldsymbol{W}_m[\boldsymbol{m}_i,\boldsymbol{X}_i]+\boldsymbol{b}_m) \tag{11}$$

$$\boldsymbol{\alpha}_g=\sigma(\boldsymbol{W}_g[\boldsymbol{g}_i,\boldsymbol{X}_i]+\boldsymbol{b}_g) \tag{12}$$

$$\boldsymbol{R}_i=\boldsymbol{\alpha}_m\otimes\boldsymbol{m}_i+\boldsymbol{\alpha}_g\otimes\boldsymbol{g}_i \tag{13}$$

式中：$\sigma(\cdot)$ 表示 Sigmoid 激活函数；⊗表示点积运算。这种加权融合方式可以使得模型更好地综合利用掩膜视觉特征与网格视觉特征，从而提高模型的视觉表征能力和描述性能。相比于传统交叉注意力机制，这种特征融合方式可以更精细地学习文本语义与掩膜特征、网格特征之间的关系，同时能够自适应地调整不同特征的重要性，有效提升模型对图像语义信息的表征能力。

将双流融合特征 $\boldsymbol{R}_i$ 输入前向反馈层形成下一个解码块的输入：

$$\boldsymbol{X}_{i+1}=\text{FeedForward}(\boldsymbol{X}_i+\text{LayerNorm}(\boldsymbol{R}_i)) \tag{14}$$

最后一层解码块的输出经过一个线性层来预测下一个单词输出 $\boldsymbol{X}_t$。

2.5 模型训练

本文模型采用两阶段训练策略，首先通过如下的交叉熵损失函数对模型进行预训练：

$$L_{\text{XE}}(\theta)=-\sum_{t=1}^{T}\log_a(p_\theta(\boldsymbol{y}_t^*\mid\boldsymbol{y}_{1:t-1}^*)) \tag{15}$$

然后，使用自评价序列训练策略在强化学习框架下训练模型，奖励函数基于 CIDEr 指标得分设计，表示如下：

$$L_{\text{R}}(\theta)=-\mathbb{E}_{\boldsymbol{y}_{1:T}\sim p_\theta}[r(\boldsymbol{y}_{1:T})] \tag{16}$$

式中：$r(\cdot)$ 表示 CIDEr 得分。

L_{R} 奖励函数的梯度优化公式表示为：

$$\nabla_\theta L_{\text{R}}(\theta)\approx -(r(\boldsymbol{y}_{1:T}^s)-r(\hat{\boldsymbol{y}}_{1:T}))\nabla_\theta\log_a p_\theta(\boldsymbol{y}_{1:T}^s) \tag{17}$$

式中：$\boldsymbol{y}_{1:T}^s$ 表示采样得到的描述语句；$\hat{\boldsymbol{y}}_{1:T}$ 表示当前模型利用贪婪策略生成的描述语句。

3 实验验证

3.1 数据集

在广泛使用的 MSCOCO 2014 数据集上评估所提方法，该数据集包括 123 287 张图像(82 783 张用于训练，40 504 张用于验证)，每张图像都有 5 个参考描述。根据通常采用的“Karpathy”划分方

法[23]，实验中重新划分了训练和测试数据集，其中113 287 张图像用于训练，5 000 张图像用于验证，5 000 张图像用于离线评估。按照惯例，将所有描述语句英文字母转换为小写表示，并删除出现次数不到 6 次的单词，最终形成了由 9 487 个单词组成的词汇表。

3.2 实验环境与超参数设置

本文实验基于 PyTorch 框架，使用的显卡为 NVIDIA RTX 3090。使用文献[22]中的 Swin-L 骨干网络提取图像网格视觉特征。使用 kMaX-DeepLab[20]全景分割网络提取图像的掩膜视觉特征。视觉特征维度设置为 512，多头注意力中的头数设置为 8，编解码 Transformer 块大小设置为 3。在训练阶段，首先在批大小为 10 的条件下使用交叉熵损失预训练模型，迭代次数设置为 20。在模型训练阶段，本实验利用 Adam 优化算法和 warmup 学习率预热技巧来优化模型，预热步数设置为 1 000。随后，利用 SCST 自评价强化学习策略继续对模型迭代训练 30 次，学习率设置为 5×10^{-6}，在模型测试阶段，本实验采用集束搜索策略，束宽设置为 3。

3.3 评价指标

本文中的图像描述生成实验使用机器翻译评价指标 CIDEr[24]、BLEU[25]、METEOR[26] 以及 ROUGE-L[27]对生成的描述进行评价，下面介绍常用评价指标计算公式。

CIDEr 是图像描述任务的评价标准，通过计算 n-gram 的 TF-IDF 来衡量生成文本和人工标注文本之间的相似性。CIDEr 计算公式如下：

$$\mathrm{CIDEr}_n(\boldsymbol{c},\boldsymbol{S})=\frac{1}{M}\sum_{i=1}^{M}\frac{\boldsymbol{g}^n(\boldsymbol{c})\cdot\boldsymbol{g}^n(\boldsymbol{S}_i)}{\|\boldsymbol{g}^n(\boldsymbol{c})\|\times\|\boldsymbol{g}^n(\boldsymbol{S}_i)\|} \tag{18}$$

$$\mathrm{CIDEr}(\boldsymbol{c},\boldsymbol{S})=\sum_{n=1}^{N}W_n\cdot\mathrm{CIDEr}_n(\boldsymbol{c},\boldsymbol{S}) \tag{19}$$

式中：$\boldsymbol{c}$ 为生成的文本；$\boldsymbol{S}$ 为参考文本集合；M 为参考文本的数量；$\boldsymbol{g}^n(\cdot)$ 为基于 n-gram 的 TF-IDF 向量；W_n 为 n-gram 的权重。

BLEU 的计算公式如下：

$$V_{\mathrm{BLEU}}=V_{\mathrm{BP}}\exp\left(\sum_{n=1}^{N}w_n\log_a p_n\right) \tag{20}$$

$$V_{\mathrm{BP}}=\begin{cases}1, & c\geqslant r\\ \mathrm{e}^{(1-r/c)}, & c<r\end{cases} \tag{21}$$

式中：c 表示生成句子的长度；r 表示参考句子的长度；p_n 表示不同的 n-gram 精度；n-gram 表示 n 长度的单词的词组集合。N 可取 1～4，分别对应 BLEU-1～BLEU-4 这 4 个评价指标，w_n 一般为所有 n 的倒数。

METEOR 是在 BLEU 的基础上得出的，同时考虑了整个语料库上的准确率和召回率。METEOR 的计算公式如下：

$$F_{\mathrm{mean}}=\frac{P_mR_m}{\alpha P_m+(1-\alpha)R_m},\ V_{\mathrm{Pen}}=\gamma\left(\frac{ch}{m}\right)^{\lambda} \tag{22}$$

$$P_m=\frac{|m|}{\sum_k h_k(c_i)},\ R_m=\frac{|m|}{\sum_k h_k(\boldsymbol{s}_{ij})} \tag{23}$$

$$V_{\mathrm{METEOR}}=(1-V_{\mathrm{Pen}})_{\mathrm{mean}} \tag{24}$$

式中：α、γ 和 λ 均为默认参数；m 为预设的一组校准；c_i 为候选语句；$\boldsymbol{s}_{ij}$ 为参考译文；$h_k(c_i)$ 表示 W_k 在候选语句中出现的次数；$h_k(\boldsymbol{s}_{ij})$ 表示 W_k 在参考译文中出现的次数；W_k 为第 k 组可能的 n-gram。

ROUGE-L 是自动文本摘要与机器翻译中常用的指标，主要通过比较生成文本与参考文本来计算两者之间的相似度。ROUGE-L 具体计算如下所示：

$$R_{ks}=\frac{\mathrm{LCS}(X,Y)}{m},P_{ks}=\frac{\mathrm{LCS}(X,Y)}{n}$$

$$F_{ks}=\frac{(1+\beta^2)R_{ks}P_{ks}}{R_{ks}+\beta^2P_{ks}} \tag{25}$$

BLEU-4 和 CIDEr 是常用的图像描述评估指标，它们不仅考虑了单词级别的对齐，还考虑了短语级别的一致性，因此能够更全面地评估生成文本的质量。此外，其对生成描述质量的评价与人类评估结果相吻合，能够在一定程度上刻画人类对图像描述结果的评价。

3.4 实验结果分析

3.4.1 定量对比分析

表 1 给出了在 MSCOCO 数据集“Karpathy”划分方式下各方法生成描述语句的性能结果，其中，B-N、M、R、C 和 S 分别是 BLEU-N、METEOR、ROUGE-L、CIDEr 和 SPICE 的缩写，比较模型包括 ORT[7]、M2 Transformer[10]、X-Transformer[11]、RSTNet[28]、Dual Global[12]、DLCT[13]、PureT[29]（包括 PureT-standard 和 PureT-Swin）、RAS-FSG[30]。ORT、X-Transformer、M2 Transformer 使用 Faster R-CNN 提取的区域特征进行实验，RSTNet 使用网格特征，PureT-standard 使用标准 Transformer 作为视觉编码器，而 PureT-Swin 使用 Swin-Transformer 作为视觉编码器。Dual Global 和 DLCT 将网格功能与区域功能相结合。RAS-

FSG[30]提出关系感知选择模块，并以语义信息作为监督知识来对网格特征进行噪声过滤。可以看到，本文所提模型的 CIDEr 指标达到了最高的 138.5，与目前最优的模型 PureT-Swin 相比，本文模型获得了更好的 BLEU-1、BLEU-4、ROUGE-L 和 CIDEr 指标，METEOR 和 SPICE 分数略低于 PureT-Swin。实验结果表明，所提方法使用全景分割掩膜特征作为视觉特征，能够有效提升图像描述生成的准确度，总体上超过了现有基于网格特征和对象区域特征的方法。

表 1 模型性能对比结果

Table 1 Comparison results of models performance

模型	B-1	B-4	M	R	C	S
ORT	80.5	38.6	28.7	58.4	128.3	22.6
X-Transformer	80.9	39.7	29.5	59.1	132.8	23.4
M2 Transformer	80.8	39.1	29.2	58.6	131.2	22.6
RSTNet	81.8	40.1	29.8	59.5	135.6	23.3
Dual Global	81.3	40.3	29.2	59.4	132.4	23.3
DLCT	81.4	39.8	29.5	59.1	133.8	23.0
RAS-FSG	81.3	40.0	29.6	59.2	134.3	23.0
PureT-standard	82.0	40.3	29.9	59.9	137.5	23.8
PureT-Swin	82.1	40.9	30.2	60.1	138.2	24.2
本文模型	82.5	41.0	30.1	60.2	138.5	24.1

3.4.2 定性对比分析

为了进一步分析全景分割掩膜特征的有效性，图 4 对标准 Transformer 模型和本文方法生成的描述语句进行了对比。其中，标准 Transformer 模型使用了 Faster R-CNN 提取的目标特征。为便于对比，真实的单词用黑体标记，本文方法生成的单词用斜体标记。可以看出，所提方法对第一行图片进行了全景分割，能够精确区分图中的 3 个年轻人目标，因此有助于进行准确的人数统计和生成正确的单词。而标准 Transformer 模型则不能利用分割目标特征生成准确的量词。在第二行图片中，“蛋糕”和“叉子”对象之间存在重叠问题，标准 Transformer 模型出现了视觉混乱和错误的关联，导致生成了不正确的单词“叉子”。相比之下，本文方法利用了全景分割模型为“勺子”提供了清晰的视觉轮廓，从而生成了正确的单词“勺子”。定性实验表明，本文方法能够有效利用全景分割掩膜特征捕捉更全面和更细粒度的视觉信息，有效缓解了视觉混乱问题。

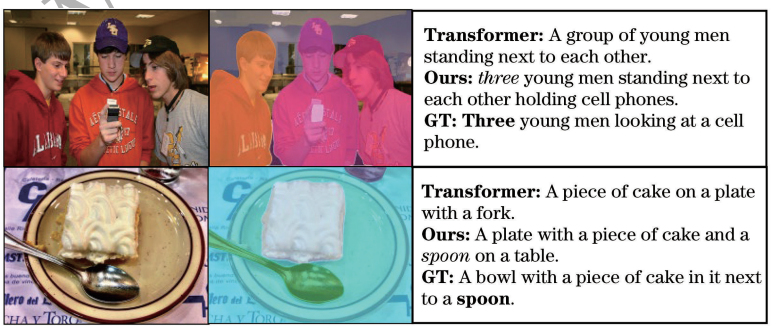
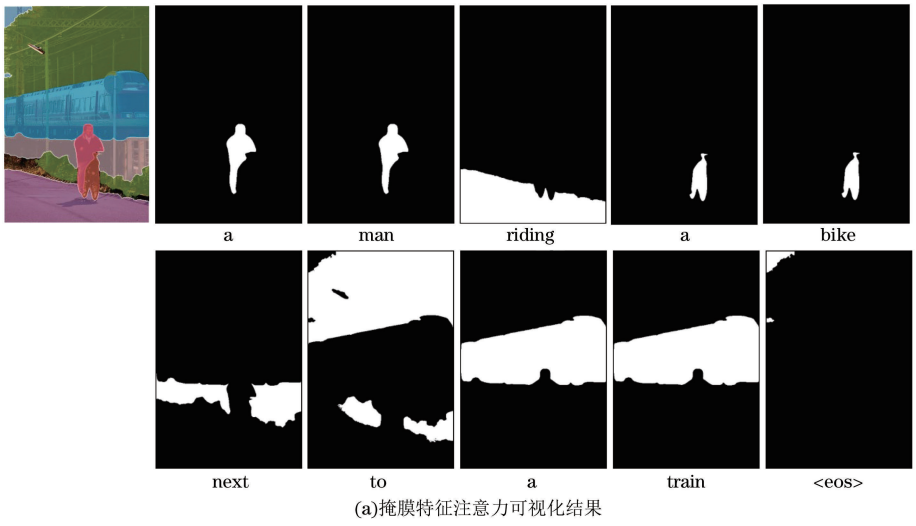


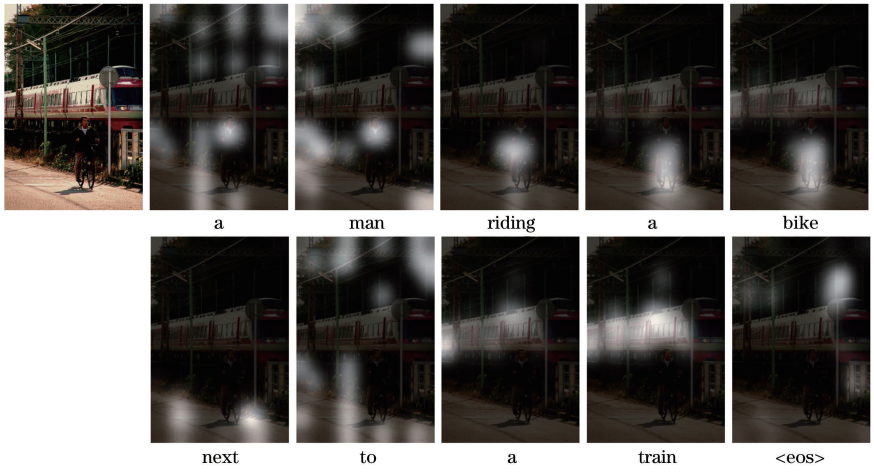
图 4 标准 Transformer 模型和本文模型生成的描述对比

Fig. 4 Comparison of descriptions generated by the standard Transformer model and our model

为了更好地评估本文模型生成单词过程的可解释性，实验中计算了最后一个解码器层的注意力权重。图 5 给出了描述生成过程中掩码特征和网格特征的注意力权重可视化结果。可以发现，本文模型在大多数情况下进行单词推断时都能关注图像的相关区域。例如，当生成“man”、“bike”和“train”等名词时，掩码注意力和网格注意力都准确地关注到与单词相关的图像区域，掩码注意力提供了清晰的轮廓，网格注意力关注上下文相关区域，2 种特征及注意力机制能够实现信息互补，有助于准确地进行描述生成。实验结果进一步验证了所提模型具有较好的可解释性。



(a)掩膜特征注意力可视化结果



(b)网格特征注意力可视化结果

图 5 基于掩膜视觉特征和网格特征的注意力权重

Fig. 5 Attention weights based on mask visual features and grid features

3.4.3 消融实验

在消融实验中，分别对 ResNet101、ResNet152 和 Swin-Transformer 提取的网格特征进行测试，以验证所提方法的有效性。如表 2 所示，R101 和 R152 分别表示 ResNet101 和 ResNet152 网络，括号中数字 7 和 14 分别表示 7×7 和 14×14 的网格特征尺寸。Swin 表示 Swin-Transformer 骨干网络，M 表示全景分割的掩膜特征。可以看出，掩膜视觉特征能够显著提升基于网格特征的模型性能。在 R101(7)、R101(14)、R152(7)和 R152(14)这 4 种不同的骨干网络上，引入掩膜视觉特征之后，CIDEr 准确性指标分别提高了 7.4、9.4、8.5 和

表 2 掩膜视觉特征与不同网格特征融合条件下的模型性能

Table 2 Models performance under the fusion conditions of mask visual features and different grid features

模型	B-1	B-4	M	R	C	S
R101(7)	78.6	36.7	28.0	57.1	123.4	22.0
R101(7)+M	81.1(2.5↑)	39.3(2.6↑)	28.9(0.9↑)	58.7(1.6↑)	130.8(7.4↑)	22.8(0.8↑)
R101(14)	78.5	36.5	28.0	57.0	121.3	21.8
R101(14)+M	81.0(2.5↑)	39.0(2.5↑)	29.0(1.0↑)	58.6(1.6↑)	130.7(9.4↑)	22.9(1.1↑)
R152(7)	78.7	36.8	28.2	57.3	123.0	22.0
R152(7)+M	81.4(2.7↑)	39.5(2.7↑)	29.0(0.8↑)	58.9(1.6↑)	131.5(8.5↑)	23.0(1.0↑)
R152(14)	78.0	36.4	28.3	57.2	123.3	22.3
R152(14)+M	81.3(3.3↑)	39.9(3.5↑)	29.2(0.9↑)	59.0(1.8↑)	132.8(9.5↑)	23.1(0.8↑)
Swin	81.6	39.8	29.9	59.6	136.4	23.8
Swin+M	82.5(0.9↑)	41.0(1.2↑)	30.1(0.2↑)	60.2(0.6↑)	138.5(2.1↑)	24.1(0.3↑)

9.5。值得注意的是，尽管 Swin-Transformer 能够提取当前最优的网格特征，但本文方法能够进一步结合掩膜视觉特征，将 CIDEr 指标得分提升 2.1。实验结果进一步证实了本文所提取的全景掩膜视觉特征与网格特征融合的有效性。

为了与当前主流图像描述方法进行公平对比，本文模型仍然使用 Swin-Transformer 作为骨干网络提取网格视觉特征。为了进一步比较目标检测特征和全景分割特征之间的差异，使用 Faster R-CNN 提取的目标特征和本文方法提取的掩码特征进行描述性能对比。如表 3 所示（其中 F 表示 Faster R-CNN 提取的区域特征），本文提取的掩膜特征无论是在单一特征还是与网格特征融合的情况下性能都优于目标区域特征，证明了其在特征解耦上的优势。

表 3 掩膜视觉特征与目标区域特征对描述性能的影响

Table 3 The influence of mask visual features and target area features on descriptive performance

模型	B-1	B-4	M	R	C	S
F	79.8	38.4	28.6	58.4	128.6	22.6
M	80.1	38.2	28.9	58.4	129.0	22.8
R101(7)+F	80.4	39.0	28.9	58.7	129.3	22.8
R101(7)+M	81.1	39.3	28.9	58.7	130.8	22.8
R101(14)+F	80.6	39.1	28.8	58.7	129.3	22.6
R101(14)+M	81.0	39.0	29.0	58.6	130.7	22.9

4 结束语

针对现有图像描述生成模型中区域视觉特征的局限性，本文提出一种新的多视觉特征融合的图像描述生成方法。该方法不仅可以有效解耦视觉表征，而且能够充分结合掩膜视觉特征和网格视觉特征的优势，提升图像描述生成的性能。在 MSCOCO 标准数据集上的定量和定性实验结果表明，所提出的方法能够显著提升现有模型的性能，同时描述生成过程具有较强的可解释性。然而，本文方法在建模掩膜视觉特征之间的层次关系和多样化描述生成能力上存在局限性，在下一步工作中，将引入场景图和扩散模型，进一步表征掩膜视觉特征的全局与局部层次关系，并结合扩散模型理论增强多样化图像描述生成的能力。

参考文献

[1] GHANDI T, POURREZA H, MAHYAR H. Deep learning approaches on image captioning: a review[J]. ACM Computing Surveys, 2024, 56(3): 1-39.

[2] LI Y P, ZHANG X R, CHENG X N, et al. Learning consensus-aware semantic knowledge for remote sensing image captioning[J]. Pattern Recognition, 2024, 145: 109893.

[3] 石义乐，杨文忠，杜慧祥，等. 基于深度学习的图像描述综述[J]. 电子学报，2021，49(10)：2048-2060.
SHI Y L, YANG W Z, DU H X, et al. Overview of image captions based on deep learning[J]. Acta Electronica Sinica, 2021, 49(10): 2048-2060. (in Chinese)

[4] STEFANINI M, CORNIA M, BARALDI L, et al. From show to tell: a survey on deep learning-based image captioning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 539-559.

[5] WANG J, XU W, WANG Q, et al. On distinctive image captioning via comparing and reweighting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2088-2103.

[6] YANG X, ZHANG H, CAI J. Deconfounded image captioning: a causal retrospect[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 12996-13010.

[7] HERDADE S, KAPPELER A, BOAKYE K, et al. Image captioning: transforming objects into words[EB/OL]. [2023-12-05]. https://arxiv.org/abs/1906.05963.

[8] GUO L T, LIU J, ZHU X X, et al. Normalized and geometry-aware self-attention network for image captioning[EB/OL]. [2023-12-05]. https://arxiv.org/abs/2003.08897.

[9] LI G, ZHU L C, LIU P, et al. Entangled transformer for image captioning[EB/OL]. [2023-12-05]. https://openaccess.thecvf.com/content_ICCV_2019/papers/Li_Entangled_Transformer_for_Image_Captioning_ICCV_2019_paper.pdf.

[10] CORNIA M, STEFANINI M, BARALDI L, et al. Meshed-memory transformer for image captioning[EB/OL]. [2023-12-05]. https://arxiv.org/abs/1912.08226.

[11] PAN Y W, YAO T, LI Y H, et al. X-linear attention networks for image captioning[EB/OL]. [2023-12-05]. https://arxiv.org/abs/2003.14080.

[12] JI J Y, LUO Y P, SUN X S, et al. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(2): 1655-1663.

[13] LUO Y P, JI J Y, SUN X S, et al. Dual-level collaborative transformer for image captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(3): 2286-2293.

[14] 李志欣，魏海洋，黄飞成，等. 结合视觉特征和场景语义的图像描述生成[J]. 计算机学报，2020，43(9)：1624-1640.
LI Z X, WEI H Y, HUANG F C, et al. Combine visual features and scene semantics for image captioning[J]. Chinese Journal of Computers, 2020, 43(9): 1624-1640. (in Chinese)

[15] 周东明，张灿龙，李志欣，等. 基于多层级视觉融合的图像描述模型[J]. 电子学报，2021，49(7)：1286-1290.
ZHOU D M, ZHANG C L, LI Z X, et al. Image captioning model based on multi-level visual fusion[J]. Acta Electronica

Sinica, 2021, 49(7): 1286-1290. (in Chinese)
[16] 刘茂福,施琦,聂礼强. 基于视觉关联与上下文双注意力的图像描述生成方法[J]. 软件学报, 2022, 33(9): 3210-3222.
LIU M F, SHI Q, NIE L Q. Image captioning based on visual relevance and context dual attention[J]. Journal of Software, 2022, 33(9): 3210-3222. (in Chinese)
[17] 宋井宽,曾鹏鹏,顾嘉扬,等. 基于视觉区域聚合与双向协作的端到端图像描述生成[J]. 软件学报, 2022, 34(5): 2152-2169.
SONG J K, ZENG P P, GU J Y, et al. End-to-end image captioning via visual region aggregation and dual-level collaboration[J]. Journal of Software, 2022, 34(5): 2152-2169. (in Chinese)
[18] CHENG B W, SCHWING A G, KIRILLOV A. Per-pixel classification is not all you need for semantic segmentation [EB/OL]. [2023-12-05]. http://arxiv.org/abs/2107.06278v2.
[19] CHENG B W, MISRA I, SCHWING A G, et al. Masked-attention mask transformer for universal image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 1290-1299.
[20] YU Q H, WANG H Y, QIAO S Y, et al. K-means mask transformer[EB/OL]. [2023-12-05]. https://link.springer.com/content/pdf/10.1007/978-3-031-19818-2_17.pdf?pdf=inline%20link.
[21] WU M R, ZHANG X Y, SUN X S, et al. DIFNet: boosting visual information flow for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2022: 18020-18029.
[22] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2021: 9992-10002.
[23] KARPATHY A, JOULIN A, LI F F. Deep fragment embeddings for bidirectional image sentence mapping[J]. Advances in Neural Information Processing Systems, 2014, 3: 1889-1897.
[24] VEDANTAM R, ZITNICK C L, PARIKH D. CIDEr: consensus-based image description evaluation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 4566-4575.
[25] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. [S. l.]: ACL, 2001: 311-318.
[26] BANERJEE S, LAVIE A. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments[EB/OL]. [2023-12-05]. https://aclanthology.org/W05-0909.pdf.
[27] LIN C Y. ROUGE: a package for automatic evaluation of summaries[EB/OL]. [2023-12-05]. https://aclanthology.org/W04-1013.pdf.
[28] ZHANG X Y, SUN X S, LUO Y P, et al. RSTNet: captioning with adaptive attention on visual and non-visual words[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2021: 15465-15474.
[29] WANG Y Y, XU J G, SUN Y F. End-to-end transformer based model for image captioning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2585-2594.
[30] LI Y N, MA Y W, ZHOU Y Y, et al. Semantic-guided selective representation for image captioning[J]. IEEE Access, 2023, 11: 14500-14510.

编辑 吴云芳