Computer Science and Application 计算机科学与应用, 2025, 15(10), 251-265
Published Online October 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1510265

Hans汉斯

基于掩码图像建模的遮挡图像分类

樊晓曼，赵其鲁*，付庆龙

青岛大学计算机科学技术学院，山东 青岛

收稿日期：2025年9月15日；录用日期：2025年10月20日；发布日期：2025年10月28日

摘　要

图像大面积遮挡导致的局部信息缺失与语义混淆是图像分类领域长期存在的挑战。为了提高遮挡图像分类准确率，本文提出一种新颖且鲁棒的基于掩码图像建模的遮挡图像分类框架(SMIM-Net)，旨在通过语义感知的掩码建模策略增强模型对遮挡区域的推理能力。该框架引入实例分割模型提取语义边界精确的语义区域作为掩码基本单元，随后通过随机掩码策略构造语义缺失的上下文，并借助通过无监督聚类算法构建的视觉词典提供高层语义监督，迫使模型基于未掩码区域推理被遮挡语义内容。在Pascal与MS-COCO数据集上的实验表明：SMIM-Net在遮挡下的平均分类准确率分别较基线提升15.7%和10.9%；在重度遮挡场景(60%~80%)下，对真实物体片段遮挡(Pascal-o)与真实遮挡(MS-COCO)的分类准确率分别达到90.8%与88.7%，领先最优方法1.5%与3.8%，为遮挡鲁棒分类提供了新范式。

关键词

遮挡图像分类，掩码图像建模，图像分割，视觉词典

Occluded Image Classification Based on Masked Image Modeling

Xiaoman Fan, Qilu Zhao*, Qinglong Fu

College of Computer Science and Technology, Qingdao University, Qingdao Shandong

Received: September 15, 2025; accepted: October 20, 2025; published: October 28, 2025

Abstract

Large-area occlusion in images leading to localized information loss and semantic ambiguity has long been a challenge in the field of image classification. To improve the classification accuracy of occluded images, this paper proposes a novel and robust occluded image classification framework based on

*通讯作者。

文章引用: 樊晓曼, 赵其鲁, 付庆龙. 基于掩码图像建模的遮挡图像分类[J]. 计算机科学与应用, 2025, 15(10): 251-265.
DOI: 10.12677/csa.2025.1510265

masked image modeling (SMIM-Net), which aims to enhance the model's reasoning capability for occluded regions through a semantic-aware masking strategy. The framework introduces an instance segmentation model to extract semantically precise boundaries as basic masking units. Subsequently, a random masking strategy is employed to construct contexts with semantic missing, while a visual dictionary built via an unsupervised clustering algorithm provides high-level semantic supervision, forcing the model to learn to infer occluded semantic content based on unmasked regions. Experiments on the Pascal and MS-COCO datasets demonstrate that SMIM-Net improves the average classification accuracy under occlusion by 15.7% and 10.9%, respectively, compared to the baseline. Under severe occlusion scenarios (60%~80%), the classification accuracy for real object segment occlusion (Pascal-o) and real occlusion (MS-COCO) reaches 90.8% and 88.7%, outperforming the best existing methods by 1.5% and 3.8%, respectively. This work provides a new paradigm for occlusion-robust classification.

Keywords

Occluded Image Classification, Masked Image Modeling, Image Segmentation, Visual Vocabulary

1. 引言

遮挡图像分类是计算机视觉的关键研究方向，在智能安防、自动驾驶等场景应用广泛，但因目标物体局部信息缺失而极具挑战。核心难点在于：遮挡的随机性与多样性阻碍有效特征提取；遮挡物与目标的语义混淆易致模型误判；合成与真实遮挡的分布差异限制了模型语义推理能力。近年来，深度学习技术蓬勃发展，计算机视觉领域涌现出多种强大的模型用于图像处理。卷积神经网络(Convolutional Neural Networks, CNN)是早期广泛应用的一类模型，它采用局部感受野逐层聚合特征的方式来提取图像信息。然而，其局部感受野特性在面对遮挡导致的局部信息缺失时，难以区分噪声和有效信号，影响分类性能。与此同时，基于注意力机制的 ViT (Vision Transformer) [1]作为一种新兴模型，在图像分类领域崭露头角。ViT 具备建模全局依赖关系的能力，能够从整体上把握图像的特征，但是其全局注意力机制在严重遮挡场景下易受无关区域干扰，导致局部判别性特征提取不足。由此可见，无论是传统的 CNN 还是新兴的 ViT，尽管在传统图像处理中都取得了显著成功，但在处理遮挡图像时都存在一定的局限性。

针对遮挡图像分类，现有工作主要围绕两个方向展开：一是通过对抗训练或者合成手段生成遮挡样本进行数据增强，旨在提升模型对遮挡图像的鲁棒性，但此类方法依赖于人工设计的遮挡模式，难以覆盖真实场景的复杂性；二是采用显式检测与修复策略，通过定位遮挡区域并重建被遮挡区域的特征。这类方法计算开销相对较大且检测精度直接影响最终性能。尽管这些方法取得了一定进展，但它们在真实遮挡的适应性、计算效率与性能稳定性上仍面临挑战。

近年来，掩码图像建模(Masked Image Modeling, MIM)思想的兴起为解决遮挡问题提供了新的视角。MIM 迫使模型学习基于上下文推理被遮挡区域信息的能力，本质上模拟了人类“补全”缺失信息的认知过程，天然具备处理局部信息缺失的潜力，为解决遮挡导致的局部信息缺失问题提供了关键思路。掩码图像建模范式自提出以来迅速发展，其中语义级 MIM 研究(如 BEiT v2 [2])通过引入知识蒸馏，利用预训练教师模型进一步将 MIM 的重建目标从像素级提升至语义级，极大地增强了预训练模型的语义表征能力。然而，BEiT v2 仍未突破传统 MIM 的核心局限：其 token 的生成依赖固定尺寸的网格分块，这种规

则的处理方式会割裂物体的自然语义边界，导致语义单元与真实物体结构脱节，并且难以适应遮挡位置、形状和尺度的多样性。为了从根本上解决上述问题，更好地将 MIM 的推理能力和遮挡场景的特性结合，本研究提出一种基于掩码图像建模的遮挡图像分类框架(Semantic-Masked Image Modeling Network, SMIM-Net)，该框架的核心创新在于首次将 MIM 的上下文推理能力针对性地迁移至遮挡场景，构架了一种具有完整语义的分割区域掩码序列建模策略，该策略通过实例分割模型 SAM (Segment Anything Model) [3]获取图像中的多个完整语义区域，并采用网格化空间定位方法依据分割区域质心的空间位置分布，将它们排列成一个结构化的输入序列。对该序列进行随机掩码后，输入基于 Transformer 的编码器并结合视觉词典进行上下文推理。具体而言，为充分发挥 MIM 在遮挡图像上的语义推理能力，本研究引入 SAM 生成语义完整的 token 表示。SAM 作为一个通用的图像分割基础模型，通过训练可以学习到极强的分割先验知识。利用 SAM 对输入图像进行处理和分析，可以将图像解构为一系列具有精确空间边界的语义分割区域。我们将这些语义分割区域作为掩码图像建模中推理的基本单元，以此改善传统 MIM 方法和现有语义级 MIM 方法的固有缺陷。在此基础上，经 k-means 无监督聚类算法聚合大量分割区域特征，构建一个覆盖广泛视觉概念的视觉词典(Visual Vocabulary)，该词典为每个分割区域赋予唯一的类别索引，并将其作为后续掩码区域语义推理的监督信号。此时掩码图像建模任务被重新定义为预测被掩码区域在预构建的视觉词典中所对应的类别索引，而不再是预测被掩码网格块的像素值或底层视觉 token。最终，利用上述预训练的编码器提取的图像特征，完成遮挡图像分类任务。

研究的创新点包括：

1) 提出基于掩码图像建模的遮挡图像分类新范式：本研究首次将掩码图像建模思想引入遮挡图像分类任务，通过设计“序列随机掩码 - 上下文推理 - 序列补全”的序列预测机制，引导模型主动学习遮挡场景下的全局语义关联规律，从根本上提升模型对目标部分可见场景的语义理解能力。

2) 基于语义分割区域的对象化序列构建机制：摒弃使用传统规则网格块构建序列的做法，创新性地引入实例分割模型生成边界精准的语义分割区域，并通过网格化空间定位方法依据区域间的空间拓扑关系将离散的语义区域排列为结构化对象序列。不仅有效保持了物体结构的完整性和空间信息，还克服了网格单元割裂对象、难以适应多样化遮挡形状的固有缺陷。

3) 视觉词典引导的类别索引预测机制：区别于 BEiT v2 等依赖有监督教师模型提供语义标签的方法，本方法通过无监督聚类构建视觉词典，以视觉词典的类别标签为监督信号，迫使模型学习分割区域间的语义依赖关系，而非单纯还原视觉细节。这种语义级预测机制增强了模型对遮挡区域的语义推理鲁棒性。

本文的结构安排如下：第 2 节对遮挡图像分类领域的研究现状展开系统性概述，着重阐述该领域内的最新研究进展，并剖析当前研究所面临的关键挑战。第 3 节针对所提出的方法进行深入探究，详细阐释模型的架构设计以及训练流程。第 4 节详细介绍本文的实验设计环节，涵盖所采用数据集的具体描述、评估指标的选取依据，以及对模型在遮挡图像分类任务中的性能表现进行全面且细致的分析。最后在第 5 节对结果进行讨论。

2. 相关工作

2.1. 掩码图像建模

掩码图像建模思想源于自然语言处理中成功的掩码语言建模(Masked Language Modeling, MLM) [4]，其核心是通过随机掩码输入的一部分，训练模型基于剩余上下文预测被掩码的内容。这一范式因其能高效引导模型学习数据的内在结构和深层表征，已成为自监督视觉表征学习的重要方向。随着 MIM 在视觉领域的兴起，一系列工作探索了不同的掩码目标和重建目标。早期研究如 SimMIM [5]和 Masked

Autoencoders (MAE) [6]采用简单的像素值重建，验证了 MIM 范式的有效性。然而，像素级重建可能导致模型过于关注局部纹理而非高层语义。为克服此局限，BEiT [7]离散变分自编码器预生成视觉词表，将重建目标从像素提升至离散的视觉标记，引导模型学习更具语义的中间表示。尽管 BEiT 方法取得了显著成功，但其基本单元为规则划分的图像块，这种预处理方式会割裂物体的自然语义边界，导致模型学习到的上下文关系建立在非语义对齐的单元上。针对这一问题，语义级 MIM 研究应运而生，旨在将重建目标进一步提升至语义空间。BEiT v2 [2]通过引入知识蒸馏，利用预训练的教师模型为被掩码区域生成软标签作为重建目标，从而将语义知识注入 MIM 过程。类似地，PeCo [8]和 MVP [9]等方法也探索了利用视觉–语言模型或对比学习来获取更优的语义监督信号。这些方法通过语义对齐的重建目标，显著增强了预训练模型的语义表征能力。然而，现有的语义级 MIM 方法仍存在核心局限。首先，其操作的基本单元依然是规则网格图像块，语义监督信号被赋予这些非语义完整的单元，存在“语义错配”问题。其次，这类方法通常依赖强大的预训练教师模型来提供监督，这引入了对额外数据或计算资源的需求，且教师模型的能力上限可能制约学生模型的发展。

2.2. 遮挡图像分类

遮挡图像处理研究围绕如何消除遮挡干扰展开，主要分为生成式遮挡修复与判别式特征抗噪两类范式。自 Goodfellow 等人[10]提出对抗生成网络(Generative Adversarial Network, GAN)以来，生成式方法已成为遮挡修复的核心技术之一。生成式遮挡修复方法依托生成模型重建被遮挡内容：早期工作如 Yu 等人[11]采用基于 GAN 的两阶段对抗生成框架，通过特征相似性度量动态对齐未遮挡区域与遮挡区域的纹理块，从而遮挡区域重建；随着 Transformer 的兴起，Yang 等人[12]设计层次化解码器，在低分辨率阶段建模全局语义关系生成主体结构。此类方法虽能生成视觉连贯结果，但计算成本较高且对严重遮挡的语义推理存在局限。判别式特征抗噪方法则通过特征解耦规避遮挡干扰：He 等人[13]利用 Transformer 对齐全局与局部语义部件，通过交叉注意力对齐可见区域与遮挡区域的上下文关系，Wang 等人[14]在 PVT 中采用稀疏注意力抑制遮挡区域响应；Cen 等人[15]提出深度特征增强(DFA)生成伪特征扩展遮挡模式覆盖，而 Yang 等人[16]通过多尺度对比学习强化遥感场景下的判别性特征。针对遮挡图像分类任务，研究者提出专项优化策略，Kotwal 等人[17]设计潜在增强网络(LEARN)，通过重构损失与类内/类间潜在空间约束提升特征可分性，强化遮挡模式下的类内紧凑性与类间可分性；Kortylewski 等人[18]结合 DCNN 与组合模型，构建物体部件的空间分布字典，动态选择组合模型或 DCNN 分支以应对不同遮挡强度。在注意力机制创新中，TDAPNet [19]通过递归自上而下注意力调节，联合原型学习与部分匹配逐步聚焦判别性区域；OSAT [20]集成视觉 Transformer 与遮挡掩码预测器(OMP)，利用自注意力权重动态抑制低置信度区域的特征响应。CompositionalNets [21]融合概率组合模型与 CNN，通过可微分 von-Mises-Fisher 分布建模特征空间实现遮挡区域概率定位；Zhao 等人[22]提出 RLA 模型，采用多尺度空间 LSTM (Long Short-Term Memory network)编码器提取遮挡鲁棒特征，并利用双通道 LSTM 解码器结合对抗训练保留身份语义。当前技术呈现生成–判别协同进化趋势：恢复方法为分类提供数据增强(如 CutMix [23])，而判别模型通过特征解耦降低对恢复精度的依赖，但计算效率与分类鲁棒性的平衡仍是亟待突破的瓶颈。

3. 方法

本文提出了一个面向遮挡图像分类的鲁棒性框架，具体流程如图 1 所示，核心思想是通过语义感知掩码图像建模预训练策略构建语义区域级的视觉理解能力，增强模型对遮挡区域的推理能力，最终应用于遮挡图像分类任务。框架包含三个关键模块：视觉语义词典的构建、分割区域掩码图像建模预训练、基于掩码图像建模预训练的遮挡图像分类。

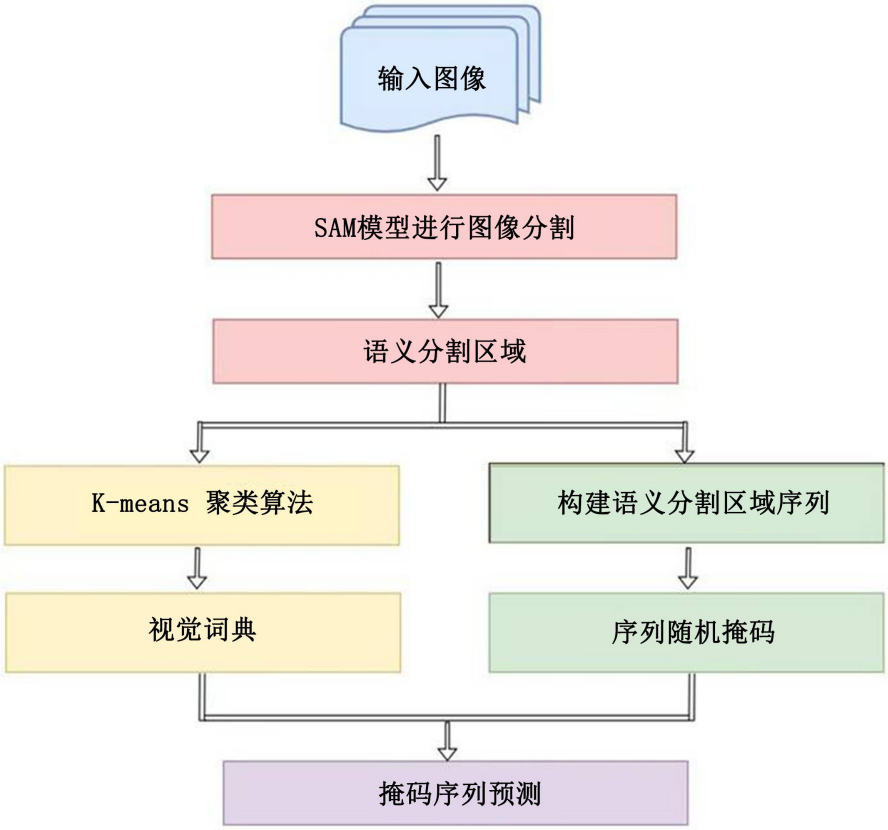

Figure 1. Schematic diagram of the training module for the occluded image classification framework SMIM-Net
图 1. 遮挡图像分类框架 SMIM-Net 训练模块示意图

3.1. 视觉语义词典构建

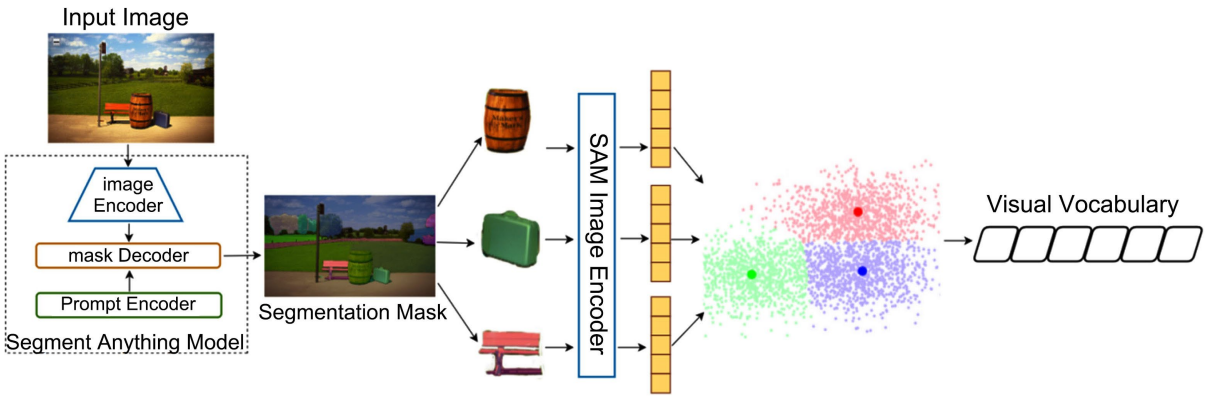

Figure 2. Schematic diagram of visual-semantic dictionary construction in the SMIM-Net framework
图 2. SMIM-Net 框架视觉语义词典构建示意图

视觉语义词典建构的核心是通过 SAM 分割和 k-means 聚类算法构建一个具有语义一致性的视觉词汇表，其示意图如图 2 所示。传统的视觉词典通常使用基于图像块的离散视觉标记，但这些标记可能缺乏明确的语义边界，例如一个图像块可能覆盖物体的一部分或背景，导致语义模糊。而本方法构建的视觉词典由 K 个聚类中心构成，每个聚类中心代表一类具有相似视觉特征的分割区域，形成泛化的语义单

元，每个分割区域则被赋予其所属语义单元在视觉词典中的索引作为类别标签。这种方式不仅更加符合视觉认知规律，又使得模型能够学习到图像的共性语义模式。

分割区域特征提取：在预处理阶段，对于给定输入图像集 $I=\{I_1,I_2,\cdots,I_Q\}$，首先采用 SAM 对每张输入图像进行分割，生成语义连贯的分割掩码集合 $M_k=\{M_{k1},M_{k2},\cdots,M_{kt}\}$，其中 $M_{kt}\in\{0,1\}^{H\times W}$。由于通过 SAM 得到的分割区域的形状和大小不一，为得到统一维度的特征表示，本方法使用 SAM 内置的基于 ViT 的图像编码器从原始图像中生成特征图，再结合分割掩码对该特征图上的对应区域进行特征提取和空间聚合，获得每个不规则分割区域的固定维度特征表示。具体而言，该编码器将原始输入图像 $I_k\in R^{H\times W\times 3}$ 分割为 16 × 16 的非重叠块，并通过多层 Transformer 块进行特征变换，最终得到对应的特征图 $F_{orig}\in R^{h\times w\times D}$，其中 $h=\frac{H}{16}$，$w=\frac{W}{16}$，D 为特征维度。此时图像编码器输出的特征图分辨率与 SAM 生成的二值掩码 M_{kt} 分辨率不同，这种分辨率差异导致无法直接在特征图上精确定位分割区域，因此需采用双线性插值函数将二值掩码下采样至特征图尺度：

$$M_{down}=\text{BilinearInterp}\left(M_{kt},h,w\right) \tag{1}$$

其中，M_{kt} 表示第 k 张输入图像的第 t 个分割区域。

此时得到的二值掩码 M_{dowm} 大小为 $h\times w$，是 $[0,1]$ 的连续概率值，表示某位置属于分割区域的置信度。为减少背景噪声的干扰，本方法随后通过阈值二值化保留分割掩码的离散属性：

$$M_{down}^{binary}[i,j]=\begin{cases}1, \text{ if } M_{down}[i,j]\geq 0.5\\ 0, \text{otherwise}\end{cases} \tag{2}$$

最终得到与特征图空间对齐的二值掩码 $M_{down}^{binary}\in\{0,1\}^{h\times w}$。为得到准确的分割位置，我们将有效分割区域定义为：

$$M=\left\{(i,j)\middle|M_{down}^{binary}[i,j]=1\right\} \tag{3}$$

每个索引 (i,j) 对应特征图 F_{orig} 中的一个空间位置。对于每个 $(i,j)\in M$，提取特征图 F_{orig} 中对应位置的特征向量构成分割区域特征向量 $F\in R^{m\times D}$，其中 $m=|M|$ 为分割区域像素数。对分割区域特征向量 F 沿着空间维度进行全局平均池化得到 $F_{pool}\in R^D$，将原本形状为 $m\times D$ 的特征压缩为一个固定长度的 D 维向量，保证了所有分割区域拥有统一的特征表示。

视觉词典生成：在预处理阶段的基础上聚合所有图像的分割区域特征，构成分割特征聚合矩阵 $F_{mask}\in R^{N\times D}$，其中 $N=\sum_{k=1}^{Q}|M_k|$ 表示所有图像分割区域总数，Q 为图像总数。以该矩阵作为输入，通过 k-means 聚类算法生成 K 个聚类中心 $\{c_1,c_2,\cdots,c_k\}$。k-means 是经典的无监督聚类算法，其目标是通过最小化特征向量到聚类中心的欧氏距离平方和将 N 个特征向量划分为 K 个互斥子集：

$$\min_{\{c_k\}_{k=1}^{K}}\sum_{k=1}^{K}\sum_{f\in c_k}\left\|f-c_k\right\|_2^2 \tag{4}$$

其中，c_k 表示第 k 个聚类中心的特征向量集合。

通过该算法每个聚类中心代表一个语义类别单元，最终生成包含 K 个语义类别的视觉词典 V，其中 K 为预设定的类别数。构建的视觉词典如图 3 所示。

这一过程本质上实现了对图像语义内容的层次化抽象，将原始像素空间映射到语义概念空间，其作用可类比于 BEiT [7]中离散变分自编码器(Discrete Variational Autoencoder, dVAE)生成的视觉标记，但通过引入语义级分割先验，能够更精准地捕获与真实语义边界对齐的视觉基元。

Figure 3. Visualization of partial categories of the visual dictionary
图 3. 视觉词典部分类别可视化

3.2. 分割区域掩码图像建模预训练

本节旨在利用掩码图像建模思想构建模型对语义分割区域上下文的理解能力。过程如图 4 所示。

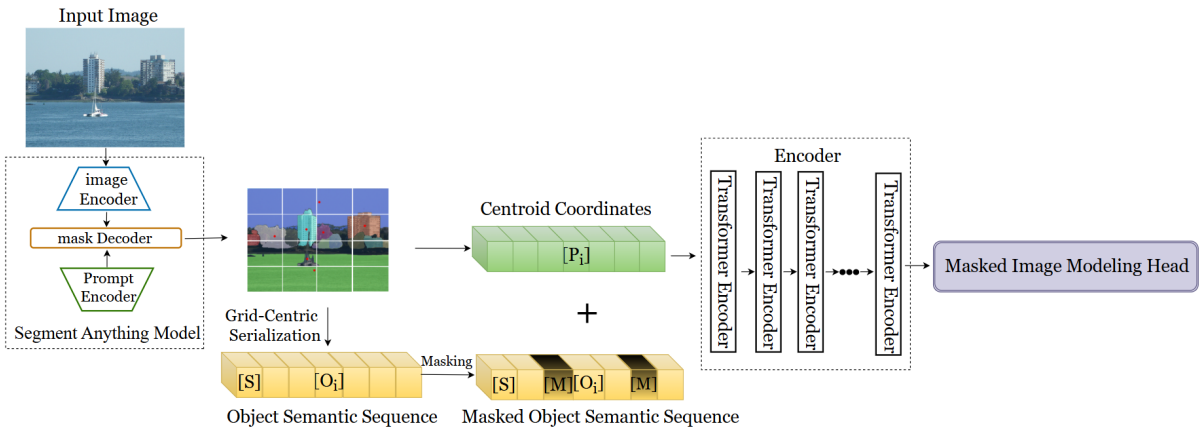

Figure 4. Schematic diagram of pre-training for segmentation region mask modeling in the SMIM-Net framework
图 4. SMIM-Net 框架分割区域掩码建模预训练示意图

构建序列化输入：传统的 Tranformer 模型(如 ViT、BEiT)通过固定网格划分破坏物体整体性。为了改善这个缺陷，本方法通过 SAM 生成语义完整的分割区域，以分割区域为基本单位构建特征序列。对于给定图像 $I \in R^{H\times W\times 3}$，利用 SAM 生成语义连贯的分割区域集合 $M=\{M_1,M_2,\cdots,M_N\}$，并通过预处理提取每个分割块的标准化特征向量 $f_i \in R^D$，其中 N 表示单张输入图像生成的语义连贯区域的数量。为了构建一个既保留物体空间拓扑关系的特征序列，本方法依据分割物体的质心坐标确定其顺序。具体地，将输入图像重塑为非重叠的 HW/P^2 个图像块，每个网格单元的分辨率为 (P,P)。计算分割块质心坐标 (x_i,y_i)，并映射至对应的网格 (P_x,P_y)。按行优先顺序遍历网格单元，若当前网格内包含一个分割物体的质心，则

将该质心对应的分割区域特征 f_i 加入序列，构成语义对象序列 $S=[f_1,f_2,\cdots,f_N]\in R^{N\times D}$。若该网格内存在多个质心，则依据这些质心在该网格单元内的位置坐标，仍按行优先顺序将其对应的分割区域特征依次加入语义对象序列。

随机掩码策略：该策略的主要目的是对语义对象序列实施随机掩码，通过人为制造的信息缺失模拟真实场景中的遮挡，迫使编码器依赖序列中可见部分预测被掩码部分的高层语义类别，从而使得每个语义单元学习与其他单元的关联特征。这一过程直接对应于遮挡图像分类的核心挑战，即当目标物体的部分区域被遮挡时，模型需要基于其他可见对象及其空间关系来推断被遮挡区域的语义属性。具体而言，在训练过程中，从$[0,N]$中均匀随机采样掩码数量 k_{mask} （k_{mask} 为整数），其中 N 为序列长度。在语义对象序列 S 中随机选择 k_{mask} 个位置生成掩码位置集合 M_{pos}，被掩码的位置使用可学习的掩码标记 $[MASK]\in R^D$ 替换该对象特征，得到掩码后的语义对象序列 S_{mask}。

$$S_{mask}=\left\{f_i \middle| f_i=\begin{cases}[MASK] & \text{if } i\in M_{pos}\\ f_i & \text{else}\end{cases}\right\} \tag{5}$$

掩码后的语义对象序列 S_{mask} 将被输入基于多层 Transformer 的编码器进行上下文推理，为弥补 Transformer 自注意力机制对位置信息的无偏性，将序列 S_{mask} 输入编码器前需与位置信息结合以保留空间信息。在本方法中位置编码 $P_i\in R^D$ 根据质心坐标生成，通过可学习的二维正弦编码映射实现：

$$P_i=\text{Linear}\left(\text{Concat}\left(\sin x_i,\cos x_i,\sin y_i,\cos y_i\right)\right) \tag{6}$$

其中，(x_i,y_i) 为第 i 个分割物体的质心所在坐标。

最终构建的特征序列为 $H^0=S_{mask}+P$，其中 $P\in R^{N\times D}$ 为位置编码矩阵。

掩码预测：不同于传统方法生成的离散局部视觉标记(Visual Token)提供监督信号，本方法利用第一阶段自监督聚类生成的视觉词典 V 作为掩码预测监督信号，将掩码预测任务提升至全局语义分类。这一设计直接驱动模型学习与高层语义类别的对齐，同时利用聚类生成的紧凑伪标签进一步减少计算开销。基于 3.1 节内容通过聚类生成的视觉词典 V，对于每个分割块特征 $f_i\in R^D$ 执行最近邻搜索，计算该特征与所有聚类中心 c_k 的欧氏距离，选择距离最小的中心索引作为该对象的伪标签 y_i：

$$y_i=\arg\min_k\left\|f_i-c_k\right\|_2^2 \tag{7}$$

将 H^0 输入 L 层 Transformer 结构进行编码，$H^l=\text{Transformer}\left(H^{l-1}\right)$，其中 $l=1,\cdots,L$。最后一层的输出向量 $H^L=\left[h_1^L;h_2^L;\cdots;h_N^L\right]$ 用作输入序列的编码表示，其中 h_i^L 是第 i 个分割区域的向量。对于每个掩码位置 $p\in M_{pos}$ 提取其编码表示 $h_p^L\in R^D$，使用 softmax 分类器预测其对应视觉词典中的类别伪标签 y_i'：

$$p_{MIM}\left(y_i'\middle|S_{mask}\right)=\text{softmax}_{y_i'}\left(W_c h_i^L+b_c\right) \tag{8}$$

其中，$W_c\in R^{|V|\times D}$，$b_c\in R^{|V|}$。在预训练中，掩码预测目标是最大化在给定被掩码序列的情况下正确对象类别 y_i 的对数似然：

$$\max\sum_{x\in D}E_M\left[\sum_{i\in M}\log p_{MIM}\left(y_i\middle|S_{mask}\right)\right] \tag{9}$$

其中，D 为训练数据，M_{pos} 代表随机掩码位置集合。

3.3. 基于掩码图像建模预训练的遮挡图像分类

经过基于视觉词典和掩码图像建模思想预训练得到的 SMIM-Net 编码器具备了从可见区域推理遮挡

区域语义的能力，为了执行最终的遮挡图像分类任务，对于输入图像首先经过预处理阶段得到每个分割区域的标准化特征向量，并依据这些分割区域质心的空间位置分布(如 3.2 节所述)，将其组织成结构化的输入序列 S，此时输入的序列是完整的、未经掩码的。将完整的序列 S 输入经过预训练的 SMIM-Net 编码器，将编码器中最后一层 Transformer 输出的编码向量作为最终分类器的输入：

$$X=\left[h_1^L;h_2^L;\cdots;h_N^L\right]\in R^{N\times D_{in}} \tag{10}$$

其中，h_i^L 是第 i 个分割区域的编码向量，D_{in} 表示输入特征维度。这些特征既包含了区域自身的语义信息，也包含了该区域与其他分割区域在上下文语义层面的依赖关系和空间位置上的关联。

为了对整个图像进行分类，采用全局平均池化(GAP)聚合区域特征，生成图像级表示：

$$h_{global}=\frac{1}{N}\sum_{i=1}^{N}h_i^{(L)} \tag{11}$$

将聚合特征输入轻量级 softmax 分类头：

$$\hat{y}=\mathrm{Softmax}\left(W_c h_{global}+b_c\right) \tag{12}$$

其中，W_c、b_c 为可学习参数，输出 $\hat{y}$ 为类别概率分布。

训练过程中采用交叉熵损失优化：

$$\mathcal{L}_{cls}=-\sum_{c=1}^{c}y_c\log\left(\hat{y}_c\right) \tag{13}$$

其中，y_c 为真实标签。

4. 实验

为了全面评估所提出的 SMIM-Net 框架在遮挡图像分类任务上的有效性，我们设计并执行了详尽的实验。本节首先介绍使用的数据集和评估指标，接着详细说明实现细节，最后系统地呈现实验结果并对结果进行多维分析。

4.1. 数据集和评估指标

数据集：实验选用了两个广泛应用于遮挡图像分类的基准数据集：Occluded-Vehicles 数据集(以下简称 Pascal)和 Occluded-COCO Vehicles 数据集(以下简称 MS-COCO)。Occluded-Vehicles 数据集是 PASCAL3D+数据集的一个子集，在[24]中提出并在[18]中扩展的。该数据集包含来自 PASCAL3D+数据集的车辆图像及其对应的分割掩码，对该数据集使用四种类型遮挡物：纯白矩形块(w)、随机像素噪声块(n)、纹理块(t)、真实物体片段(o)。遮挡程度根据遮挡面积占比可分为四个等级：L0 (0%)、L1 (20%~40%)、L2 (40%~60%)、L3 (60%~80%)。图 5 展示了不同遮挡等级和类型的示例：(a) 干净图像 0%遮挡；(b) 随机噪声和白色噪声构成的 20%~40%遮挡；(c) 自然物体构成的 40%~60%遮挡；(d) 纹理构成的 60%~80%遮挡。Pascal 数据集中的遮挡是合成的，通过测试人工生成的遮挡来评估框架是合理的，但是为了研究框架在真实遮挡情况下的有效性，我们也考虑在包含真实遮挡的 MS-COCO 数据集上评估所提出的 SMIM-Net 框架。MS-COCO 数据集根据图像中车辆目标的遮挡比例同样划分为 L0~L3 四个遮挡等级，每个遮挡等级的测试图像数量分别为：2036 张(L0)、768 张(L1)、306 张(L2)、73 张(L3)。其核心价值在于评估模型在真实、不可控、自然发生的遮挡场景下的泛化能力。MS-COCO 中的遮挡来源于真实世界的复杂交互，其遮挡物的形状、纹理、语义关系均不可预测且高度多样化。

评估指标：为全面衡量模型性能，我们采用分类准确率(%)作为主要评估标准，反映模型预测的最可

能类别等于真实类别的图像比例，直接表征模型的整体分类能力。为了专门评估模型对被遮挡区域语义推理的能力(这是 SMIM-Net 第二阶段预训练的核心目标)，我们使用遮挡语义召回率(Occlusion Semantic Recall, OSR, %)指标，该指标计算模型在预测图像整体类别的同时，正确预测出被遮挡物体区域在视觉词典 V 中对应类别索引的比例(仅针对被掩码的区域进行评估)，较高的 OSR 值表明模型能有效利用上下文信息推理遮挡内容并与高层语义建立关联。

(a) (b) (c) (d)

Figure 5. Examples of clean and occluded images from the Occluded-Vehicles dataset
图 5. 来自 Occluded-Vehicles 数据集的干净图像和被遮挡图像的示例

4.2. 实现细节

我们的实验严格遵循 SMIM-Net 的三阶段流程进行，所有的代码基于 PyTorch 1.12 实现，在 8 × NVIDIA A100 GPU (80 GB)上进行训练和评估。

视觉词典构建：基于 ImageNet-1K 训练集(1.28 M 图像)，采用官方 SAM-ViT-H 模型进行实例分割，利用 SAM 图像编码器提取每个分割区域的 1024 维特征向量。对约 1200 万个分割区域特征进行 k-means 聚类(k = 8000)，生成视觉词典，过程中使用 Faiss 库加速。

分割区域掩码图像建模预训练：基于 ImageNet-1K 训练集(1.28 M 图像)，骨干网络采用 12 层 Transformer 架构(隐藏维度为 768，注意力头数 12)。共训练 100 个 epoch，权重衰减 0.05，批大小 1024，学习率采用分阶段衰减策略：第 1~25 个 epoch 初始学习率为 0.01，第 25~50 个 epoch 学习率降至 0.001，第 50~100 个 epoch 学习率进一步衰减至 0.0001。

下游任务的微调：在 4.1 节所述的数据集上进行评估，加载预训练编码器，接入一个轻量级的 softmax 分类器。模型训练优化采用 SGD，学习率 0.1，动量 0.9，权重衰减 1e−4，批大小 128，共训练 35 个 epoch。

我们与多种代表性基线进行对比，为公平比较，所有基线模型均采用相同输入分辨率(224 × 224)并在相同下游数据集微调。

4.3. 实验结果

4.3.1. Pascal 数据集上的结果

我们在 Pascal 数据集上评估了 SMIM-Net 在不同遮挡程度和不同遮挡类型下的性能，并与一些最新的最先进方法进行了比较，如 CompositionalNets [21]、基于字典的组合模型[18]、TDAPNet [19]和 LATENT [17]。其中遮挡程度分为 L0：0%、L1：20%~40%、L2：40%~60%、L3：60%~80%四个等级；遮挡类型包括：纯白色矩形块(w)、随机像素噪声块(n)、从其他自然图像中裁剪的纹理块(t)、利用 SAM 分割的其他自然物体片段(o)。同时计算了各方法在所有遮挡等级和类型上的平均准确率(%)。实验结果如表 1 所示。

可以观察到，SMIM-Net 在所有遮挡等级(L1~L3)和所有遮挡类型上均展现出显著优势，平均准确率高达 95.9%，较最优模型 CompNet (95.4%)提升了 0.5%，优于基线模型 BEiT-Base (80.2%)，提升高达 15.7%。

这充分证明了 SMIM-Net 框架在应对合成遮挡方面的整体有效性。在中重度遮挡场景(L2/L3)及语义复杂的真实物体遮挡(o 型)场景下，SMIM-Net 的优势尤为突出：在遮挡率达 60%~80% (L3)时，SMIM-Net 对真实物体片段遮挡(o)的分类准确率达 90.8%，分别领先 CompNet (88.4%)和 LEARN (89.3%) 2.4%和 1.5%。这直接证实了基于 SAM 结构语义单元并通过视觉词典改善传统掩码图像建模的有效性。值得注意的是，标准 BEiT-Base 模型在遮挡下性能急剧下降(L3 平均仅 60.6%)，凸显了传统网格级 MIM 在遮挡鲁棒性上的根本局限。SMIM-Net 通过语义单元替代网格单元以及语义类别预测替代视觉 token 预测，实现了较大的提升。

Table 1. Occluded image classification performance under varying occlusion levels and different occlusion types on the Pascal dataset
表 1. Pascal 数据集上不同遮挡程度和不同遮挡类型下的遮挡图像分类性能

Occ. Area	L0: 0%	L1: 20%~40%				L2: 40%~60%				L3: 60%~80%				Mean
Occ. Type	-	w	n	t	o	w	n	t	o	w	n	t	o	-
BeiT-Base	97.2	94.6	93.8	93.2	89.7	86.8	84.0	85.4	80.3	70.5	66.4	63.7	60.6	80.2
CoD	92.1	92.7	92.3	91.7	92.3	87.4	89.5	88.7	80.3	70.2	80.3	76.9	87.1	87.1
TDAPNet	99.3	98.4	98.6	98.5	97.4	96.1	97.5	88.7	91.6	82.1	88.1	82.7	79.8	92.8
CompNet	99.3	98.6	98.6	98.8	97.9	98.4	98.4	97.8	94.6	91.7	90.7	86.7	88.4	95.4
LEARN	100	99.7	99.8	99.6	99.0	98.3	99.0	97.8	96.1	80.5	91.9	84.4	89.3	95.1
SMIM-Net	99.6	99.3	99.5	99.1	99.4	98.0	98.0	98.2	98.1	88.3	90.2	88.6	90.8	95.9

4.3.2. MS-COCO 数据集上的结果

为了评估 SMIM-Net 在更复杂、不可控的真实遮挡下的分类性能，我们在更具挑战性的、包含自然真实遮挡的 MS-COCO 数据集上进行了性能测试，并与一些最新的最先进方法进行了比较。在实验过程中遮挡程度同样分为 L0：0%、L1：20%~40%、L2：40%~60%、L3：60%~80%四个等级，同时计算了各方法在所有遮挡等级和类型上的平均准确率(%)。实验结果如表 2 所示。

Table 2. Occluded image classification performance across varying occlusion levels on the MS-COCO dataset
表 2. MS-COCO 数据集上不同遮挡程度下遮挡图像分类性能

Occ. Area	L0	L1	L2	L3	Mean
BEiT-Base	94.6	84.4	80.6	71.1	82.7
CoD	91.8	82.7	83.3	76.7	83.6
TDAPNet	98.0	88.5	85.0	74.0	86.4
CompNet	98.5	93.8	87.6	79.5	89.9
LEARN	99.2	93.3	91.1	84.9	92.1
SMIM-Net	98.9	94.2	93.1	88.7	93.6

通过实验结果我们可以看到，在真实遮挡场景的 MS-COCO 数据集上，SMIM-Net 进一步验证了其泛化能力。SMIM-Net 以 93.6%的平均准确率全面领先现有方法，较特征增强方法 LEARN (92.1%)和组合模型 CompNet (89.9%)分别提升 1.5%与 3.7%。最具挑战性的重度真实遮挡(L3, 60%~80%)场景下，SMIM-Net 达到了 88.7%的准确率，显著优于 LEARN (84.9%)和 CompNet (79.5%)，这表明构建视觉词典能够较

好地覆盖真实世界的语义多样性，且基于语义单元的掩码建模策略能成功迁移至不可控的真实遮挡环境。值得注意的是，BEiT 基线模型在真实遮挡下性能骤降(L3: 71.1%)，暴露了标准 Transformer 全局注意力机制在严重真实遮挡下易受无关区域干扰的缺陷，而 SMIM-Net 通过输入基于分割区域的对象序列并执行视觉词典监督的序列预测有效聚焦于分类关键物体的语义信息，显著提升了模型对真实世界遮挡分布的适应能力。

4.4. 消融实验

在本节中，我们进行了一系列的消融实验，旨在系统性地评估关键组件对模型最终性能的贡献，以深入了解各个因素对整体表现的影响。我们将着重关注以下两个方面：视觉词典大小对模型的影响、掩码比例对模型的影响。

4.4.1. 视觉词典大小

我们的模型的第一个阶段使用了 k-means 算法来对分割区域特征进行聚类。使用 k-means 聚类算法时需要预先指定聚类中心的数量，这个数量对应了生成的视觉词典的大小 K，每个聚类中心代表了一个类别伪标签。视觉词典大小 K 决定了分割语义单元的粒度，直接影响模型对遮挡区域的语义推理精度。我们进行了五组实验探索不同的聚类中心数量的选择 K 对模型性能的影响。

首先，计算不同 K 值预训练模型对应的遮挡语义召回率(OSR, %)，衡量模型基于上下文推理被掩码物体语义类别的能力。如图 6 所示，揭示了视觉词典大小 K 与 OSR 的非单调关系：K = 8000 时 OSR 最高(89.4%)，而 K = 2000 时语义粒度粗糙(OSR = 68.5%)，无法区分相似物体；K = 10000 时 OSR 下降至 85.9%，因过细划分引发类内方差增大。

其次，微调预训练模型在 MS-COCO 和 Pascal 数据集上评估分类准确率(%)。为精准评估视觉词典大小 K 对模型处理重度遮挡能力的影响，我们着重报告模型在最具挑战性的场景下的性能：1) Pascal L3-o：Pascal 数据集上 60%~80%遮挡率(L3)、由真实物体片段(o 型)造成的遮挡。该场景因高遮挡率和真实语义混淆最具挑战性；2) MS-COCO L3：MS-COCO 数据集上 60%~80%遮挡率(L3)的真实世界遮挡。该场景代表开放环境下不可控的复杂遮挡；3) Pascal Mean/MS-COCO Mean：各自数据集上所有遮挡等级和类型的平均准确率，反映词典大小对整体鲁棒性的影响。实验结果如表 3 所示：当 K = 2000 时，下游分类任务准确率为 64.7% (Pascal L3-o)和 60.2% (MS-COCO L3)；随着 K 增至 8000，下游性能达峰值 90.8% (Pascal L3-o)与 88.7% (MS-COCO L3)；而 K = 10000 时，下游平均准确率在两个数据集上同步降低 2.2%和 5.2%，反映过细划分导致类内一致性下降与噪声敏感。这一现象印证了视觉词典需在语义抽象与实例泛化间取得均衡，K = 8000 时聚类中心既能覆盖多样化物体模式，又能保持足够样本支撑类别表征，为分割区域语义推理提供了最优监督信号。

Table 3. Impact of different visual vocabulary sizes K on model performance
表 3. 不同视觉词典大小 K 对模型性能的影响

K	Pascal L3-o	MS-COCO L3	Pascal Mean	MS-COCO Mean
2000	64.7	60.2	67.5	62.3
4000	78.6	75.4	80.2	77.6
6000	86.2	84.5	88.7	86.0
8000	90.8	88.7	95.9	93.6
10,000	89.3	83.8	93.7	88.4

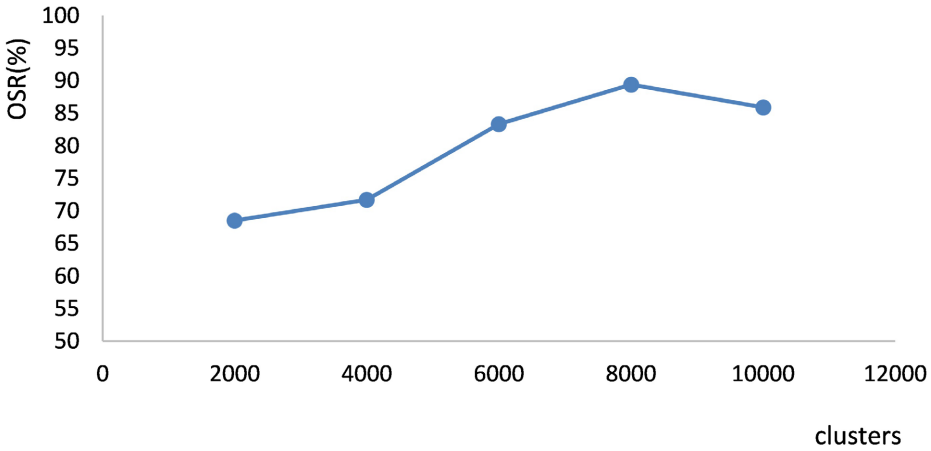

Figure 6. Impact of different visual vocabulary sizes on occluded semantic recall rate (%)
图 6. 不同视觉词典大小对遮挡语义召回率(%)的影响

4.4.2. 随机掩码策略的影响

在掩码图像建模范式中，随机掩码策略通过人为制造信息缺失以模拟真实遮挡场景，考虑到真实场景中的遮挡具有随机性，模型需要同时具备处理完整图像与遮挡图像的能力，我们设计了两种对比掩码策略：1) 完整掩码策略(Full Masking Strategy)：掩码数量 k_{mask} 从$[0, N]$中均匀随机采样，即允许零掩码的情况；2) 强制掩码策略(Forced Masking Strategy)：掩码数量 k_{mask} 从$[1, N]$中均匀随机采样，即每次至少掩码一个区域。在预训练过程中，当采样到的掩码数量为 0 时，该样本不产生掩码预测损失，模型仅进行一次前向传播。这种设计旨在让模型接触到完整的、未被破坏的图像序列，从而增强其对全局语义结构的表征学习能力。

Table 4. Impact of different masking ratios on model performance
表 4. 不同掩码比例对模型性能的影响

Masking Strategy	Pascal L0	MS-COCO L0	Pascal L3-o	MS-COCO L3	Pascal Mean	MS-COCO Mean
Full Masking	99.3	98.9	90.8	88.7	95.9	93.6
Forced Masking	97.7	96.4	87.6	83.4	93.8	80.4

在预训练阶段，两种策略均使用相同的视觉词典大小和训练配置得到预训练模型。首先在 ImageNte-1k 验证集上评估掩码预测任务的遮挡语义召回率(OSR, %)，结果如图 7 所示：采用完整掩码策略的模型 OSR 达到了 89.4%，显著高于强制掩码策略模型的 85.7%。这表明强制掩码策略因完全排除完整图像样本，导致模型在语义推理时过度依赖上下文线索，而忽视了对物体整体语义的表征学习。

进一步，我们在 Pascal 和 MS-COCO 数据集上微调两种预训练策略得到的编码器，评估其在下游遮挡图像分类任务上的性能。为全面评估不同掩码策略对下游分类任务的影响，本实验除聚焦于上述核心挑战场景(即 Pascal L3-o、MS-COCO L3、Pascal Mean、MS-COCO Mean)外，还额外计算了 Pascal L0/MS-COCO L0 完整无遮挡场景下的分类准确率，反映模型学习到的基础表征能力结果。结果如表 4 所示。对于完整无遮挡图像，完整掩码策略在 Pascal L0 (99.3%)和 MS-COCO L0 (98.9%)上分别领先强制策略 1.6%和 2.5%。这一现象进一步揭示强制掩码策略的表征退化问题：缺少完整图像训练使模型无法充分学习类别判别性特征，直接损害了基础分类能力。在重度遮挡情况下，完全掩码策略在 Pascal 和 MS-COCO 数据集上分别领先强制掩码 3.2%和 5.3%，此差异暴露了强制掩码策略在遮挡下的泛化缺陷。

总的来说，存在零掩码样本促使模型学习完整语义结构的表征能力，避免因强制遮挡导致的“局部特征依赖偏置”。

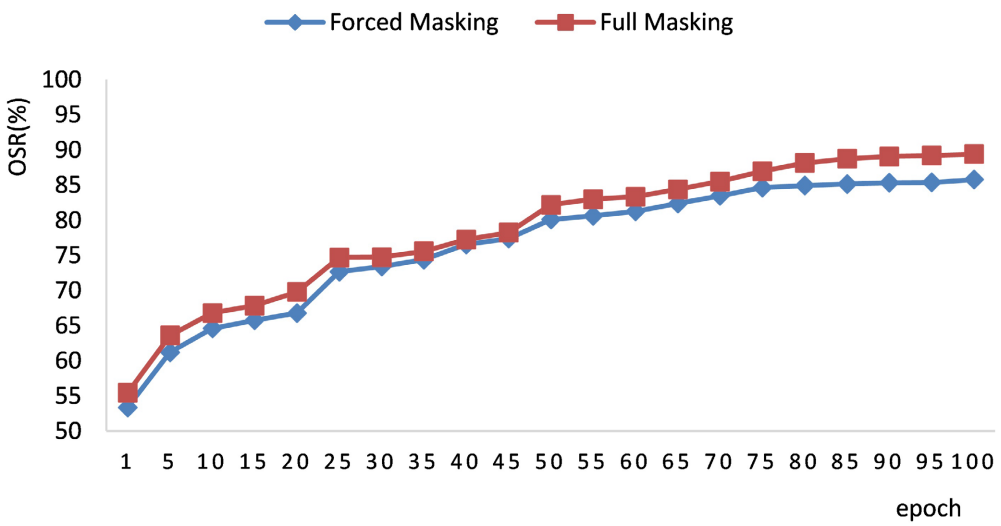

Figure 7. Impact of different masking strategies on occluded semantic recall rate (%)
图 7. 不同掩码策略对遮挡语义召回率(%)的影响

5. 结论

本研究提出并验证了一种新的遮挡图像分类框架，创新性地将掩码图像建模思想应用于遮挡图像分类任务。不同于传统的掩码图像建模方式，本框架使用具有完整语义的分割区域作为处理单元，并利用视觉词典提供高层语义监督，将掩码图像建模的重建目标从像素或低级特征提升至语义级别，迫使模型学习上下文关联而非表面纹理。实验结果表明，该框架在合成遮挡(Pascal)和真实遮挡(MS-COCO)场景下均展现出显著优势，尤其在重度遮挡及语义复杂的真实物体遮挡场景下。这充分验证了所提出的基于语义单元的掩码建模预训练策略能够有效适配真实世界遮挡的多样性与不可控性，为复杂环境下的遮挡图像分类提供了可靠的技术支撑。

基金项目

山东省自然科学基金委员会，省自然科学基金面上项目，ZR2023MF106，基于特征融合的多任务自监督视觉学习。

参考文献

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., *et al*. (2020) An Image Is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. arXiv: 2010.11929.

[2] Peng, Z., Dong, L., Bao, H., *et al*. (2022) Beit v2: Masked Image Modeling with Vector-Quantized Visual Tokenizers. arXiv: 2208.06366.

[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., *et al*. (2023) Segment Anything. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 4015-4026. https://doi.org/10.1109/iccv51070.2023.00371

[4] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Vol. 1, 4171-4186.

[5] Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., *et al*. (2022) SimMIM: A Simple Framework for Masked Image

Modeling. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 9653-9663. https://doi.org/10.1109/cvpr52688.2022.00943

[6] He, K., Chen, X., Xie, S., Li, Y., Dollar, P. and Girshick, R. (2022) Masked Autoencoders Are Scalable Vision Learners. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, 18-24 June 2022, 16000-16009. https://doi.org/10.1109/cvpr52688.2022.01553

[7] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., *et al*. (2023) Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 19175-19186. https://doi.org/10.1109/cvpr52729.2023.01838

[8] Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., *et al*. (2023) Peco: Perceptual Codebook for BERT Pre-Training of Vision Transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 552-560. https://doi.org/10.1609/aaai.v37i1.25130

[9] Wei, L., Xie, L., Zhou, W., Li, H. and Tian, Q. (2022) MVP: Multimodality-Guided Visual Pre-Training. In: Avidan, S., *et al*., Eds., *European Conference on Computer Vision*, Springer, 337-353. https://doi.org/10.1007/978-3-031-20056-4_20

[10] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., *et al*. (2014) Generative Adversarial Nets. *Communications of the ACM*, **63**, 139-144.

[11] Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X. and Huang, T.S. (2018) Generative Image Inpainting with Contextual Attention. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 5505-5514. https://doi.org/10.1109/cvpr.2018.00577

[12] Yang, F., Yang, H., Fu, J., Lu, H. and Guo, B. (2020) Learning Texture Transformer Network for Image Super-Resolution. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 5791-5800. https://doi.org/10.1109/cvpr42600.2020.00583

[13] He, S., Luo, H., Wang, P., Wang, F., Li, H. and Jiang, W. (2021) TransReID: Transformer-Based Object Re-Identification. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 15013-15022. https://doi.org/10.1109/iccv48922.2021.01474

[14] Wang, W., Xie, E., Li, X., Fan, D., Song, K., Liang, D., *et al*. (2021) Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 548-558. https://doi.org/10.1109/iccv48922.2021.00061

[15] Cen, F., Zhao, X., Li, W. and Wang, G. (2021) Deep Feature Augmentation for Occluded Image Classification. *Pattern Recognition*, **111**, Article ID: 107737. https://doi.org/10.1016/j.patcog.2020.107737

[16] Yang, Z., Chen, J., Li, J. and Zheng, X. (2025) Multiscale Occlusion-Robust Scene Classification in Remote Sensing Images via Supervised Contrastive Learning. *IEEE Geoscience and Remote Sensing Letters*, **22**, 1-5. https://doi.org/10.1109/lgrs.2025.3537104

[17] Kotwal, K., Deshmukh, T. and Gopal, P. (2024) Latent Enhancing Autoencoder for Occluded Image Classification. 2024 *IEEE International Conference on Image Processing* (*ICIP*), Abu Dhabi, 27-30 October 2024, 894-900. https://doi.org/10.1109/icip51287.2024.10647790

[18] Kortylewski, A., Liu, Q., Wang, H., Zhang, Z. and Yuille, A. (2020) Combining Compositional Models and Deep Networks for Robust Object Classification under Occlusion. 2020 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Snowmass, 1-5 March 2020, 1322-1330. https://doi.org/10.1109/wacv45572.2020.9093560

[19] Xiao, M., Kortylewski, A., Wu, R., Qiao, S., Shen, W. and Yuille, A. (2020) TDMPNet: Prototype Network with Recurrent Top-Down Modulation for Robust Object Classification under Partial Occlusion. In: Bartoli, A. and Fusiello, A., Eds., *Computer Vision—ECCV* 2020 *Workshops*, Springer International Publishing, 447-463. https://doi.org/10.1007/978-3-030-66096-3_31

[20] Heo, J., Wang, Y. and Park, J. (2022) Occlusion-Aware Spatial Attention Transformer for Occluded Object Recognition. *Pattern Recognition Letters*, **159**, 70-76. https://doi.org/10.1016/j.patrec.2022.05.006

[21] Kortylewski, A., He, J., Liu, Q. and Yuille, A.L. (2020) Compositional Convolutional Neural Networks: A Deep Architecture with Innate Robustness to Partial Occlusion. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 8940-8949. https://doi.org/10.1109/cvpr42600.2020.00896

[22] Zhao, F., Feng, J., Zhao, J., Yang, W. and Yan, S. (2018) Robust LSTM-Autoencoders for Face De-Occlusion in the Wild. *IEEE Transactions on Image Processing*, **27**, 778-790. https://doi.org/10.1109/tip.2017.2771408

[23] Yun, S., Han, D., Chun, S., Oh, S.J., Yoo, Y. and Choe, J. (2019) Cutmix: Regularization Strategy to Train Strong Classifiers with Localizable Features. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 6023-6032. https://doi.org/10.1109/iccv.2019.00612

[24] Wang, J.Y., Zhang, Z.S., Xie, C.H., *et al*. (2015) Unsupervised Learning of Object Semantic Parts from Internal States of CNNs by Population Encoding. arXiv: 1511.06855.