文章编号：1671-7449(2024)01-0048-06

基于ResNeSt和改进Transformer的多标签图像分类算法

王　贺，张　震

（山西大学 物理电子工程学院，山西 太原 030006）

摘　要：目前，基于深度学习的多标签分类算法还存在一些问题，如标签之间的相关性有待提高，如何解决小目标分类等。为此提出了一种多标签图像分类算法，该算法使用分裂注意力网络ResNeSt进行特征提取，并使用BatchFormerV2与Transformer形成双分支网络对特征进行编码，解码阶段使用Transformer Decoder的交叉注意模块来自适应地处理特征以达到更好的分类效果。实验结果表明：该模型在COCO数据集上的mAP为88.4%，在VOC2007数据集上的平均精度为96.0%，一定程度上提高了多标签图像分类的准确率。

关键词：深度学习；多标签分类；ResNeSt；Transformer

中图分类号：TP391　　**文献标识码**：A　　**doi:** 10.3969/j.issn.1671-7449.2024.01.007

引用格式：王贺，张震．基于ResNeSt和改进Transformer的多标签图像分类算法[J]．测试技术学报，2024，38(1)：48-53.

WANG He，ZHANG Zhen. Multi-label image classification algorithm based on ResNeSt and improved transformer[J]. Journal of Test and Measurement Technology，2024，38(1)：48-53.

Multi-Label Image Classification Algorithm Based on ResNeSt and Improved Transformer

WANG He，ZHANG Zhen

（College of Physics and Electronic Engineering，Shanxi University，Taiyuan，030006，China）

Abstract：At present，there are still some problems in the multi-label classification algorithm based on deep learning，such as the relevance between labels needs to be improved，and how to solve the problem that small targets are more difficult to identify than large targets. In this paper，we propose a multi-label image classification algorithm that uses the split attention network ResNeSt for feature extraction and uses a dual-branch Transformer to query class labels. In addition，we use the cross-attention module in Transformer Decoder to extract the local features adaptively. On this basis，in order to enhance the classification effect of the Transformer module，we introduce BatchformerV2 to make the Transformer form a double-branch network. The mAP of the model on the COCO dataset is 88.4%，and the average precision on the VOC2007 dataset is 96.0%，which improves the accuracy of multi-label image classification to a certain extent.

Key words：deep learning；multi-label classification；ResNeSt；Transformer

收稿日期：2023-02-17

作者简介：王　贺(1983－)，女，博士，讲师，主要从事机器学习和信息处理研究。E-mail：wanghe@sxu.edu.cn。

0　引　言

图像分类作为计算机视觉中的基本问题备受关注，常应用于图像检索[1]、视频注释[2]和指纹识别[3]等领域。通常所研究的图像中并非只包含一类对象，为了能够更准确地提取图像信息，人们开始对多标签图像分类（Multi-Label Image Classification，MLIC）展开研究。

早期多标签图像分类将图像信息转化为多个二分类方法进行训练，效率较低。2017年，Zhu F等[4]提出了SRN对目标出现区域进行研究，利用注意图对每个标签应关注的图像区域进行划分，使多标签图像分类效率有了极大提高。随着卷积神经网络（CNN）的发展，多标签图像分类的精度逐步提升，但是图像中目标之间的关联性还有待发掘，导致多标签图像分类精度不高。2019年，Chen Z M等[5]提出了ML-GCN，为图像上出现的标签建立了相关矩阵，使卷积神经网络一定程度上学习到了标签之间的关联性，在多标签图像分类上取得了很好的结果。但是ML-GCN模型获得的标签相关性具有一定的局限性，因为在数据集中标签之间的频率关系并不代表真实场景中目标之间的关系。Chen T等[6]提出了SSGRL，利用语义解耦模块结合类别语义来学习特定语义的表示，将语义交互模块与基于统计标签共现的图相关联，并通过图传播机制探索它们之间的交互，效率获得了很大提升。针对ML-GCN存在的问题，Ye J等[7]提出的ADD-GCN在一定程度上进行了改善，该方法在图像中建立了利用注意力机制驱动的动态图卷积网络来表示标签的相关性，获得了更好的效果。

近些年，Transformer开始应用于计算机视觉的各个领域，并获得了巨大的成功，如Vision Transformers（ViT）用于图像分类[8]；DETR用于目标检测[9]。受到DETR的启发，2021年，Liu S等[10]提出利用ResNet作为特征提取网络，Transformer作为分类器进行多标签分类：通过多头注意力机制，从对象的不同部分或不同视图中提取特征，将每个标签类视为Transformer解码器中的查询，并对后续二进制分类的相关特征进行交叉关注，该方法在多个数据集上获得了很好的结果。此后，Ridnik T等[11]提出基于Transformer的分类头ML-decoder对图像进行分类，ML-decoder灵活高效，可以用于单标签、多标签和零样本等多种场景中。

本文提出了一种基于ResNeSt与双分支Transformer的网络结构。首先，ResNeSt网络在不同的网络分支上应用通道注意力，利用特征图注意力和多路径表示的互补优势，能更好地提取细微的特征，提高特征提取性能；其次，在Transformer模块基础上，加入BatchFormerV2[12]模块分支，通过密集表示学习提高小样本的类别精度。与其他多标签图像分类模型相比，本文所提多标签图像分类算法在不同数据集上的精度都有一定的提升。

1　系统框架

如图1所示，系统框架由两部分构成。第一部分进行特征提取，其中特征提取模块选用分裂注意力网络（ResNeSt），它将图像进行分割，利用通道注意力提取特征，再将其整合到一个统一的注意力块中，以提取到不同网络分支的特征。

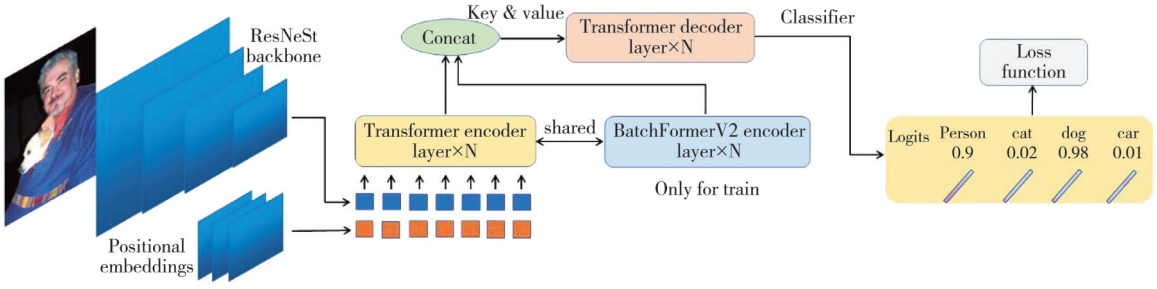

图1　系统框图

Fig. 1　Structure diagram of system

模型第二部分由Transformer Encoder、Decoder模块、BatchFormerV2 Encoder模块、分类器（Classifier）和损失函数（Loss Function）组成。将ResNeSt提取到的底层特征加入位置编码（Positional Embeddings）作为Transformer Encoder模块的输入；Transformer Encoder模块在训练阶段可以通过共享其模块得到BatchFormerV2 Encoder模块，形成双分支结构，两个分支进行

融合得到键值对传入 Transformer Decoder 模块中；然后，通过分类器进行类别预测；最后，通过损失函数进行反向传播，更新参数以得到最佳结果。需要注意的是，BatchFormerV2 Encoder 模块只在训练阶段出现，在测试阶段会被删除，这样可以避免额外的推理负担。

1.1 分裂注意力网络(ResNeSt)

ResNet解决了网络加深出现的退化问题，但是它缺乏通道间的信息融合，因此，本文采用了一种在ResNet基础上进行改进的网络——分裂注意力网络(ResNeSt)。其在ResNet的基础上加入了各个通道间的信息融合，将通道级注意力应用于不同的网络分支，在捕获跨特征交互和学习多样化表现方面具有重要作用，通过分裂注意力机制获取每个通道不同的重要性，可以将感兴趣区域更加精确地提取出来，得到比ResNet更好的效果。

ResNeSt模块如图 2 所示，输入特征图被分割为t个基础特征图，其中每个组进行切片处理得到N个切片；经过1×1卷积和3×3卷积处理之后送入分裂注意力(见图3)中；然后，将每个通过分裂注意力得到的输出进行融合操作，通过1×1卷积还原通道数，这样得到的特征融合了各个通道之间的信息，而且输入和输出的形状相同，可以当作一个模块加入模型中。

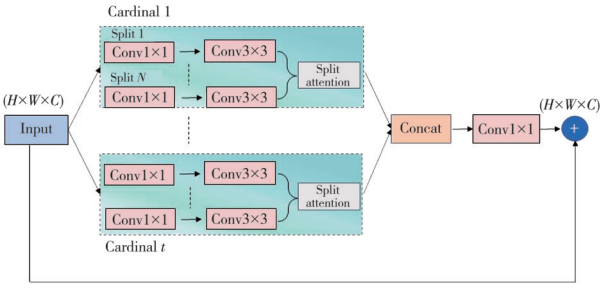

图 2 ResNeSt模块

Fig. 2 ResNeSt block

分裂注意力模块如图 3 所示，在图 2 中经过3×3卷积处理后的N个切片进入分裂注意力模块，首先，进行累加求和得到总特征，然后，通过平均池化层收集全局上下文信息，将收集到的信息通过全连接层、归一化层和ReLU函数得到每个切片的注意力权重大小，其分别与每个切片相乘再进行融合，这样可以对各个通道进行信息融合，而且可获取到每个通道的重要程度。

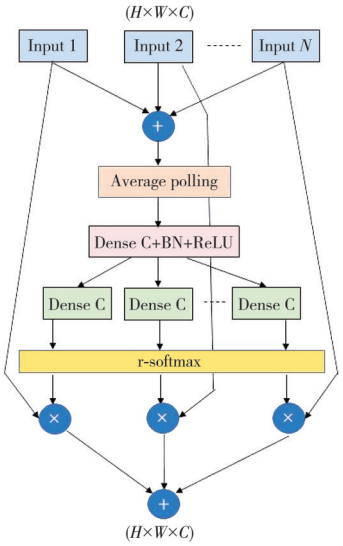

图 3 分裂注意力模块

Fig. 3 Split attention module

1.2 BatchFormerV2模块

本文引入了一种新的Transformer结构：BatchFormerV2。该模块提出了一种密集表示的样本学习方法，不仅具有空间注意，而且加入了批处理注意，实现了从图像级到像素级的表示。相比于Transformer需要大量图片进行训练而言，引入的BatchFormerV2模块可以缓解样本稀缺问题，这有利于多标签图像分类精度的提升，Transformer中注意力模块的输出Z表示为

$$Z=\mathrm{softmax}\left(\frac{\boldsymbol{QK}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}, \tag{1}$$

式中：$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}\in\mathrm{R}^{N\times C}$分别为查询矩阵、键矩阵和值矩阵，$N$为图像特征块的数量，$C$为嵌入维度；$d$为通道数。查询矩阵$\boldsymbol{Q}$与键矩阵$\boldsymbol{K}$内积，得到一个注意力矩阵，它表示了$\boldsymbol{Q}$与$\boldsymbol{K}$的相关程度，对该注意力矩阵进行归一化并通过softmax激活函数，然后点乘值矩阵$\boldsymbol{V}$得到Z。

BatchFormerV2的表达式为

$$Z_i=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V}_i^{\mathrm{T}},Z=concat(Z_1,\cdots,Z_N), \tag{2}$$

式中：$\boldsymbol{Q}_i$，$\boldsymbol{K}_i$，$\boldsymbol{V}_i\in\mathrm{R}^{B\times C}$，$Z\in\mathrm{R}^{B\times N\times C}$，$B$为Batch Size，指一个批量的数据。可以看出BatchFormerV2并没有改变Transformer的形式，只是将N个长度为B的序列视为一个批量，送入共享的Transformer模块中，其中$N=H\times W$。

如图 4 所示，将分裂注意力网络得到的底层

特征输入到Transformer Encoder模块中，在原始Transformer Encoder分支保持不变的情况下，通过共享创建一个新的分支，在这个分支中加入了BatchFormerV2模块，即两个分支共享Transformer Encoder模块。两个分支得到Key和Value后进行融合，输入到Transformer Decoder部分。Transformer Decoder部分的交叉注意力可以自适应地提取局部特征，通过查询矩阵$\boldsymbol{Q}$进行查询得到最后的结果。同时为了避免额外的参数增加，在解码阶段删除了BatchFormerV2分支。

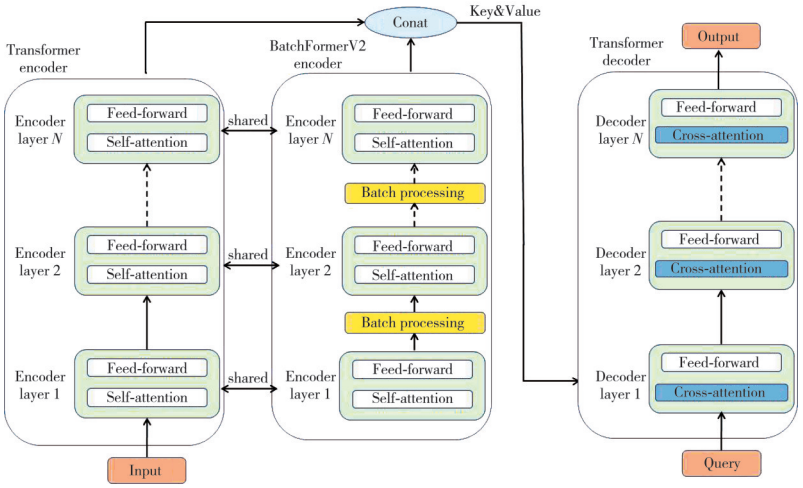

图 4　双分支Transformer模块结构图

Fig. 4　Structure diagram of dual-branch Transformer module

2　实验与分析

实验配置为：intel i7-11700CPU，64位Windows 10操作系统，Nvidia GeForce RTX 3060Ti。基于Pytorch深度学习框架搭建网络模型。使用AdamW优化器对网络优化，初始学习率调整为1×10^{-6}，并使用cutmix进行数据增强。将图像统一裁剪为448×448大小，使用均值[0,0,0]和标准差[1,1,1]对输入图像进行归一化，并使用RandAugment[13]进行增强。

为了评估提出的多标签图像分类方法，在Microsoft Common Objects in Context（MS COCO14）数据集和The PASCAL Visual Object Classes（VOC2007）数据集上进行了实验。MS COCO14包括82 783张训练集图像，40 504张验证集图像，涵盖80个常见的类别，平均每张图像上有2.9个标签，可以用于对象检测和分割，也广泛用于多标签图像分类。VOC2007数据集中包含训练集（5 011幅）和测试集（4 952幅），总共9 963幅图，涵盖了20个常见的类别。

2.1　评估标准

本文主要使用平均精度均值（Mean Average Precision，mAP）作为模型的评估指标。其中平均精度均值的计算方式为

$$AP=\frac{\sum P}{k},\tag{3}$$

$$P=\frac{TP}{TP+FP},\tag{4}$$

$$mAP=\frac{\sum_{i=1}^{N}AP_i}{N},\tag{5}$$

式中：TP（True Positive）为被判定为正样本，事实上也是正样本；FP（False Positive）为被判定为正样本，但事实上是负样本；P为准确率；k为数据集中含有该类别的图片数量；N为总类别个数；AP为单标签精度值；mAP为所有目标类别精度的平均值。

2.2　实验结果分析

本文所提出模型在MS COCO14数据集上的训练损失如图5所示，横坐标为迭代训练次数epoch，训练了25个epoch，在第20个epoch后网络趋于收敛。在MS COCO14数据集上的mAP曲线图如图6所示。本文数据都是在图像分辨率为448×448上得到的，但是为了比较的公平性，还对图像分辨率为576×576进行训练，方便与其他模型进行对比。

由表1和表2的数据对比可以发现，本文提出的模型所得到的mAP值最高，在图像分辨率为

448×448 的情况下，相比最高的 ASL 模型[14]，其 *mAP* 值高出 2.0%；在图像分辨率为 576×576 的情况下，相比最高的 Q2L-R101 模型，其 *mAP* 值高出 1.9%。说明该多标签分类算法的分类精度更高，效果更好。

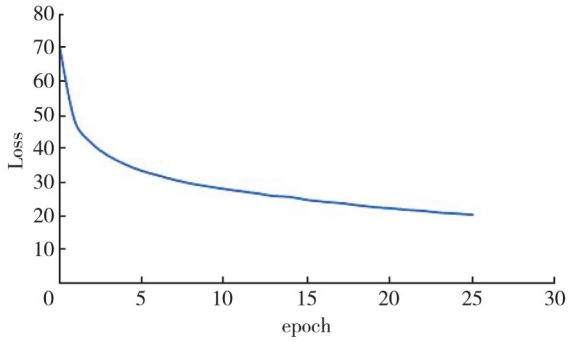

图 5 训练损失曲线

Fig. 5 Loss curve of training

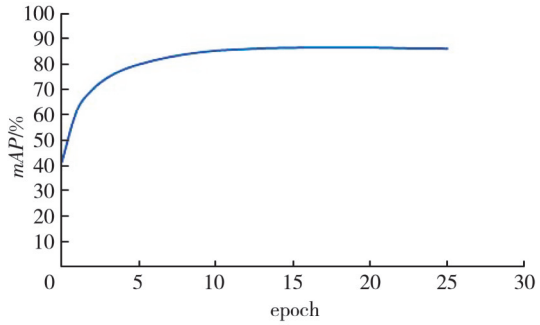

图 6 mAP 曲线

Fig. 6 Curve of mAP

表 1 各算法结果比较(448×448)

Tab. 1 Comparison of the results of each algorithm (448×448)

模型	图像分辨率	*mAP*/%
ResNet-101	448×448	78.3
ML-GCN	448×448	83.0
MCAR[15]	448×448	83.8
Q2L-R101	448×448	84.9
ASL[14]	448×448	85.0
本文改进模型	448×448	87.0

表 2 各算法结果比较(576×576)

Tab. 2 Comparison of the results of each algorithm (576×576)

模型	图像分辨率	*mAP*/%
SSGRL	576×576	83.8
C-Trans[16]	576×576	85.1
ADD-GCN	576×576	85.2
Q2L-R101	576×576	86.5
本文改进模型	576×576	88.4

另外，在 VOC2007 数据集上进行了补充实验，结果为表 3 所示。可以看出，本文所提模型平均类别精度 mAP 达到了 96.0%，在 20 个类别中有 13 个类别都有不同的提升，相比于 ResNet-101，ML-GCN，SSGRL 和 ASL 分别提高了 5.2%，2.0%，1.0% 和 0.2%。对于 ASL 只有小幅度增加是因为 VOC2007 的 mAP 值已经趋近于饱和，但是对于以前的工作仍然有很大的提升。比如，与 ASL 相比，在 bike，chair，tv 类别上分别提高了 0.7%，1.4% 和 0.8%，这说明所提方法是有效的。但是在某些类别上，所提模型的 mAP 也出现了下降，比如在 plant 类别上，比最高的 ASL 低 1.3%，故还需要进一步去研究，减少在个别类别上的差距。

表 3 各算法在 VOC2007 数据集上 *mAP* 对比

Tab. 3 Comparison of the mAP of each algorithm on VOC2007 data set

模型	*mAP*/%				
	ResNet-101	ML-GCN	SSGRL	ASL (TResNetL)	本文改进模型
aero	99.1	99.5	99.7	99.9	99.9
bike	97.3	98.5	98.4	98.4	**99.1**
bird	96.2	98.6	98.0	98.9	**99.2**
boat	94.7	98.1	97.6	98.7	98.7
bottle	68.3	80.8	85.7	86.8	86.4
bus	92.9	94.6	96.2	98.2	**98.8**
car	95.9	97.2	98.2	98.7	98.6
cat	94.6	98.2	98.8	98.5	**99.1**
chair	77.9	82.3	82.0	83.1	**84.5**
cow	89.9	95.7	98.1	98.3	**98.6**
table	85.1	86.4	89.7	89.5	89.5
dog	94.7	98.2	98.8	98.8	**99.1**
horse	96.8	98.4	98.7	99.2	99.2
mbike	94.3	96.7	97.0	98.6	**99.1**
person	98.1	99.0	99.0	99.3	**99.4**
plant	80.8	84.7	86.9	89.5	88.2
sheep	93.1	96.7	98.1	99.4	**99.7**
sofa	79.1	84.3	85.8	86.8	**87.0**
train	98.2	98.9	99.0	99.6	**99.8**
tv	91.1	93.7	93.7	95.2	**96.0**
mAP	90.8	94.0	95.0	95.8	**96.0**

2.3 消融实验

在消融实验中，进行了以下两组对比实验：1）使用 ResNet 网络和 ResNeSt 网络比较，2）使用 Transformer 模块和双分支 Transformer 模块比较。在 MS COCO14 数据集中进行测试，使用 mAP 精度值进行评价。

从表 4 中对比可以发现，减少本文算法中的任何一个模块都会导致平均精度均值 mAP 下降。当算法中都使用 Transformer 时，在 mAP 指标上 ResNeSt 网络可以比 ResNet 网络提升 1.05% 左右；当算法中都使用 ResNet 网络时，双分支 Transformer 模块会比 Transformer 模块提升 0.51% 左右；当算法中使用 ResNeSt 和双分支 Transformer 模块时，会比只

使用ResNet和Transformer模块提升1.59%左右。证明本文算法结合这两个模块可以得到更高的多标签图像分类精度。

表 4　消融实验指标对比表

Tab. 4　Comparison table of ablation experiment indicators

ResNet	√		√	
ResNeSt		√		√
Transformer	√	√		
双分支Transformer			√	√
mAP/%	85.41	86.46	85.97	87.0

3　结 论

本文提出了一种简单高效的多标签图像分类模型，该模型基于ResNeSt分裂注意力网络提取底层特征，在改进后的双分支Transformer网络上进行多标签图像分类。介绍了模型的整体框架，并说明选用ResNeSt分裂注意力网络作为骨干网络可以得到更加丰富的特征信息，使用改进后的双分支Transformer网络模型可以提高Transformer的分类效果。通过实验表明，本文所提出的多标签分类模型在mAP指标上优于其他算法。

参考文献：

[1] 曹莉华，柳伟，李国辉. 基于多种主色调的图像检索算法研究与实现[J]. 计算机研究与发展，1999，36（1）：96-100.
CAO Lihua，LIU Wei，LI Guohui. Research and implementation of an image retrieval algorithm based on multiple primary colors[J]. Journal of Computer Research and Development，1999，36(1): 96-100.（in Chinese）

[2] ALKSASBEH M Z，AL-OMARI A H，ALQARALLEH B A Y，et al. Smart hand gestures recognition using K-NN based algorithm for video annotation purposes[J]. Indonesian Journal of Electrical Engineering and Computer Science，2021，21（1）：242-252.

[3] CHEN S，GUO Z，LI X，et al. Query2Set：single-to-multiple partial fingerprint recognition based on attention mechanism[J]. IEEE Transactions on Information Forensics and Security，2022，17：1243-1253.

[4] ZHU F，LI H，OUYANG W，et al. Learning spatial regularization with image-level supervisions for multi-label image classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：5513-5522.

[5] CHEN Z M，WEI X S，WANG P，et al. Multi-label image recognition with graph convolutional networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2019：5177-5186.

[6] CHEN T，XU M，HUI X，et al. Learning semantic-specific graph representation for multi-label image recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision，2019：522-531.

[7] YE J，HE J，PENG X，et al. Attention-driven dynamic graph convolutional network for multi-label image recognition[C]//European Conference on Computer Vision，2020：649-665.

[8] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，et al. An image is worth 16x16 words：Transformers for image recognition at scale [J]. arXiv preprint arXiv：2010.11929，2020.

[9] CARION N，MASSA F，SYNNAEVE G，et al. End-to-end object detection with transformers[C]//European Conference on Computer Vision，2020：213-229.

[10] LIU S，ZHANG L，YANG X，et al. Query2label：a simple transformer way to multi-label classification [J]. arXiv preprint arXiv：2107.10834，2021.

[11] RIDNIK T，SHARIR G，BEN-COHEN A，et al. Ml-decoder：scalable and versatile classification head[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision，2023：32-41.

[12] HOU Z，YU B，WANG C，et al. BatchFormerV2：exploring sample relationships for dense representation learning[J]. arXiv preprint arXiv：2204.01254，2022.

[13] CUBUK E D，ZOPH B，SHLENS J，et al. Randaugment：practical automated data augmentation with a reduced search space [C]//Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition Workshops，2020：702-703.

[14] RIDNIK T，BEN-BARUCH E，ZAMIR N，et al. Asymmetric loss for multi-label classification [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision，2021：82-91.

[15] GAO B B，ZHOU H Y. Learning to discover multi-class attentional regions for multi-label image recognition [J]. IEEE Transactions on Image Processing，2021，30：5920-5932.

[16] LANCHANTIN J，WANG T，ORDONEZ V，et al. General multi-label image classification with transformers [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition，2021：16478-16488.