刘保虎. 融合 SAM 掩码的弱监督语义分割伪标签生成算法[J]. 智能计算机与应用,2025,15(10):156-162. DOI:10.20169/j.issn.2095-2163.251024

# 融合 SAM 掩码的弱监督语义分割伪标签生成算法

刘保虎
(上海理工大学 机械工程学院, 上海 200093)

**摘 要**: 弱监督语义分割中生成的伪标签质量对最后的分割效果起决定作用,现有的伪标签生成算法多采用 Vision Transformer(ViT)为特征提取网络,由于 ViT 提取特征过度平滑导致激活区域泛化,因此本文提出融合(Segment Anything Model, SAM)生成的伪标签来引导泛化的伪标签聚焦目标区域。首先,对原图进行数据增强,融合不同增强图像的分割伪掩码来提高 SAM 生成伪标签的精度;然后,对融合后的伪标签进行评估,使用熵值给有效像素分配更高的权重;另外,在 ViT 中插入类令牌对比模块(Class Token Contrast, CTC),该模块可以促进非显著局部对象和全局对象表示的一致性,使伪标签包含更多的目标区域;最后,通过卷积网络将 SAM 生成的伪标签和 ViT 网络生成的伪标签进行融合。实验证明,本文生成的伪标签在电芯蓝膜数据集上测试,精度达到 89.51%,平均分割率达到 85.65%。

**关键词**: 视觉转换器; SAM 引导; 类令牌对比; 伪标签优化
**中图分类号**: TP391; TP274 **文献标志码**: A **文章编号**: 2095-2163(2025)10-0156-07

# Weakly supervised semantic segmentation pseudo-label generation algorithm based on SAM mask

LIU Baohu
(School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

**Abstract**: The quality of pseudo-labels generated in weakly supervised semantic segmentation plays a crucial role in determining the final segmentation results. Existing pseudo-label generation algorithms often utilize the Vision Transformer (ViT) as the feature extraction network. However, the over-smoothing of features by ViT leads to generalized activation regions. Therefore, this paper proposes the integration of pseudo-labels generated by the Segment Anything Model (SAM) to guide the focus of generalized pseudo-labels on target regions. Firstly, data augmentation is applied to the original image, and pseudo-masks from different augmented images are fused to enhance the accuracy of SAM-generated pseudo-labels. Subsequently, the fused pseudo-labels are evaluated, and higher weights are assigned to pixels with more reliable predictions using entropy values. Additionally, a Class Token Contrast (CTC) module is introduced into ViT to promote consistency between representations of non-significant local objects and global objects, enriching pseudo-labels with more target regions. Finally, convolutional networks are employed to merge pseudo-labels generated by SAM and ViT. Experimental results demonstrate that the pseudo-labels generated in this paper, tested on the battery blue film dataset, achieve an accuracy of 89.51% and an average segmentation rate of 85.65%.

**Key words**: Vision Transformer; SAM-guided; class token contrast; pseudo-label optimization

## 0 引 言

近年来,新能源汽车逐渐取代燃油车成为人们买车的首选,新能源车的畅销使得动力电池的市场占有率得到提高。研究可知组成动力电池模组的电芯在生产过程中会包覆一层蓝膜[1],在生产中蓝膜易出现划痕、破损等缺陷,不仅影响产品外观,严重时会对动力电池安全产生危害,因此蓝膜缺陷检测对于产品质量把控具有重要意义。现阶段工业检测上应用较多的是人工目视法以及基于深度学习的有监督缺陷检测方法[2],基于深度学习的图像分割模型能准确高效地识别出蓝膜表面的各种缺陷,但模型所需数据却依赖于人工标注,不仅效率低下,而且成本过高。因此,迫切需要将对标注依赖性更低的弱监督蓝膜表面缺陷检测算法引入工业产品生产过程中[3]。

**作者简介**: 刘保虎(1998—),男,硕士研究生,主要研究方向:机器视觉。Email:443487501@qq.com。

**收稿日期**: 2024-01-17

弱监督语义分割(Weak Supervised Semantic Segmentation, WSSS)旨在通过优化和利用弱注释来满足对精确像素级语义标注的需求,尤其在仅有图像级类标签可用的情况下,这一方法的应用尤为困难。缺乏准确的位置注释使得图像级的 WSSS 方法通常需要依赖由类激活映射(Class Activation Map, CAM)生成的粗糙位置注释。CAM 是一种深度学习技术,依赖于深度分类网络来产生与各个类别相对应的特征图[4]。为了获得更加精确且高质量的伪掩码,许多既有的 WSSS 方法引入了一个细化阶段,这一阶段主要是利用更加复杂和精确的算法来改善初始的 CAM,此后会将经过细化的伪掩码用于再训练阶段,以更好地指导分割网络的学习过程。针对上述问题,本文提出了一种全新的伪标签生成方法,可以在不细化 CAM 种子的前提下改进伪标签质量。利用最近提出的 Segment Anything Model(SAM)模型,使用 SAM 增强伪标签的质量,SAM 能够在不知道类标签的情况下精确分割大部分相似区域或同一对象。因此,SAM 可以成为提高 WSSS 中伪标签质量的有力补充工具。使用 CAM 派生的伪标签 P1 与 SAM 生成的伪标签 P2,通过卷积操作选择两者最相关的片段,再对其进行标记以生成该类的新伪标签。SAM 生成的片段表现出高精度,使得现有伪标签中部分错误激活问题得到了改善。

本文主要贡献如下:

(1)将弱监督语义分割算法用于电芯蓝膜表面缺陷检测,与全监督方法相比,简化了标注方式,减少了人力物力的支出。

(2)为了获取到更为完整的蓝膜表面缺陷掩码,采用 ViT 网络生成的伪标签与 SAM 生成的伪标签融合来增强伪标签质量。

(3)在 ViT 网络特征提取过程中为了保证学习特征的一致性,研发设计了类令牌的对比模块,以促进局部非显著区域与全局对象之间的表示一致性,从而进一步强制 CAM 中激活更多的对象区域。

## 1　相关工作

### 1.1　图像级标注语义分割方法

为了从图像级标签中获取像素级标签,许多方法都专注于如何优化 CAM。SEC 方法[5]通过种子扩展传播稀疏 CAM 标签。DSRG[6]结合种子区域生长方法来扩展 CAM。DGCN[7]使用传统的图切割算法为种子周围的区域分配标签。AffinityNet[8]和 IRNet[9]使用随机游走方法传播标签,帮助 CAM 更多地关注未区分的区域。SEAM[10]探索了 CAM 在不同仿射变换下的一致性。此外,还有一些方法选择引入网络数据,例如 Co-segmentation[11]和 STC[12]。然而使用 AffinityNet[8]等方法来细化 CAM,然后使用细化的伪掩码重新训练 DeepLab[13]网络可能过于复杂且耗时。除此之外,CAM 通常只激活最具辨别力的区域,这也是 CAM 存在的主要问题与不足。为此针对这一问题,后期陆续提出了各种训练方案,例如擦除[10]、在线注意力积累[5]和交叉图像语义挖掘[14]等,都是利用辅助任务来规范训练目标,如视觉词学习[15]、子类别探索[16]和尺度不变性正则化[17]。通常,这些方法建立在 CNN 网络之上,继承了局部性缺陷,使用 ViT 框架的 WSSS 可以避免这个缺点并实现完整的对象激活,提高伪掩码的整体质量,图 1 展示了弱监督算法的一般流程。

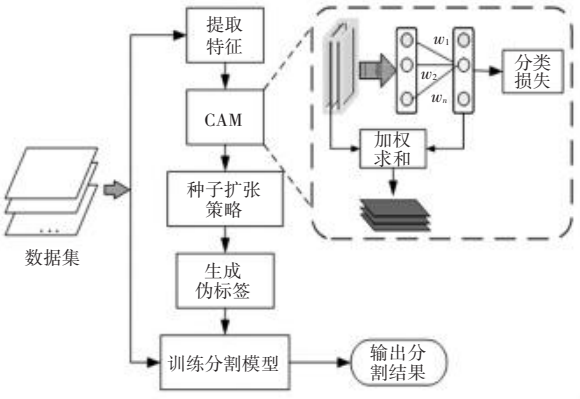


**图 1　基于 CAM 的弱监督语义分割算法一般流程**

**Fig. 1　General workflow of weakly supervised semantic segmentation algorithm based on CAM**

### 1.2　Transformer 在弱监督分割中的应用

最近 Transformer 推动了计算机视觉领域的发展。ViT 将图像转换为不重叠的补丁序列令牌[15],这些补丁序列令牌用作网络输入并添加类令牌作为输入。然后使用全连接层将类标记映射到类预测。这种没有卷积核大小限制的模型架构可以更好地挖掘图像的整体信息。TS-CAM[17]方法使用类标记和补丁序列令牌之间的交叉注意力图来获得弱监督域中的位置线索。交叉注意力图的获取需要对同一层下不同头部的注意力图进行平均,然后对不同的层求和。之后,将交叉注意力图与通过使用卷积处理补丁序列令牌获得的 CAM 相结合。TransCAM[18]是基于 CNN 和 Transformer 双主干网络,是 Transformer 和卷积的混合,还使用补丁序列注意力图在 CAM 生成阶段细化 CAM。但是,Transformer 的大多数注意力头都注意到图像中的不同位置和类

别,这包含许多与目标对象无关的信息。

### 1.3 分割任何图像模型

分割任何图像模型(SAM)视觉基础模型[19]是在超过 10 亿个 mask 的 SA1B 上训练的模型。模型的主要目标是分割任何给定图像中的任何物体,而不需要任何额外标注信息。其出色的分割效果和对新场景的零镜头泛化使 SAM 成为各种计算机视觉任务的理想候选者。在本文的蓝膜数据集中使用 SAM 来生成分割掩码是完全可以的。但是,SAM 在分割能见度较差的物体时效果不是很好,例如伪装物体、医用息肉和透明玻璃,在本文的数据集中包含的一类缺陷压痕就是属于能见度差的类型,为了进一步提高生成伪标签的准确性,提出一种伪标签细化策略。这种策略赋予那些可靠的预测更高的权重,从而得到可靠的伪标签,最终提高分割器的分割性能。

## 2 理论与方法

### 2.1 总体框架

总体框架使用 ViT-base(ViT-B)作为主干,在 CAM 的生成过程中通过类令牌对比模块来提高激活的完整性,由激活完整的 CAM 生成伪标签 $P_1$,然后与 SAM 引导生成的伪标签 $P_2$ 进行融合生成最终的伪标签 $P$。基于 SAM 引导的弱监督语义分割流程如图 2 所示。

对于 CAM 的生成过程,首先将输入图像分割成 $N^2$ 个小图像块、并将其铺平,再将其线性映射到 $N^2$ 个补丁序列令牌中。进一步生成 $C$ 个可学习的类令牌,这里 $C$ 表示分类类别的总数,并将其与补丁序列令牌连接为 Transformer 编码器的输入 $I \in R^{(C+N^2)\times D}$,其中 $D$ 为输入令牌的维数,Transformer 编码器内部由 $K$ 个编码层组成。每一层由 2 个子层组成:一个多头自注意(MSA)和一个多层感知器(MLP)。在每个子层之前进行层归一化(LN),在每个子层之后进行残差连接。在每个编码层中,输入令牌 $I$ 并接收 $O$。$O$ 成为下一个编码器层的新 $I$,以此类推 $K$ 次迭代。将图像补丁序列令牌和多个类令牌输入到 Transformer 编码器中。通过对补丁序列令牌重新排列,应用卷积来生成粗略的 CAM 记作 $M_c$,计算公式如下:

$$M_c = \sum_i W_{c,i} F_i \tag{1}$$

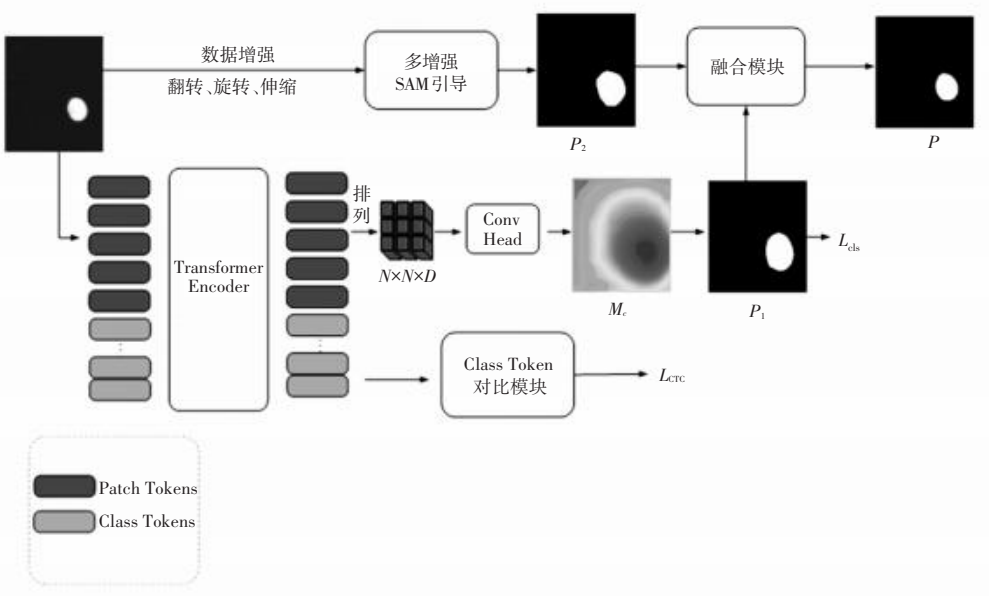


**图 2 基于 SAM 引导的弱监督语义分割流程图**

**Fig. 2 Flowchart of weakly supervised semantic segmentation guided by SAM**

### 2.2 类令牌对比模块

为了保证 CAM 激活的完整性以及学习特征的一致性,设计了类令牌的对比模块,如图 3 所示。该模块可以促进局部非显著区域与全局对象之间的表示一致性,从而进一步强制 CAM 中激活更多的对象区域,给定一张蓝膜图像,首先从其辅助 CAM 指定的非显著区域中随机裁剪局部图像。由于 ViT 中的类令牌捕获的是对象的高级语义信息,全局和局部图像的类令牌分别聚合了全局和局部对象的信息,通过最小化全局类令牌和局部类令牌之间的差异,

使得整个对象区域的表示可以更加一致。为了解决裁剪后的局部图像可能只包含很少或没有前景对象的情况，从背景区域裁剪一些局部图像，通过最大化全局图像和局部背景区域的类标记之间的差异，可以突出前景和背景的差异。这里使用 2 个背景阈值 $L$、$H$，其中 $(0 < L < H < 1)$，本文背景阈值 $(L,H)$ 设置为 $(0.25,0.70)$。将 $M_c$ 分割为可靠前景，背景和非显著区域组成的伪类令牌标签 $Y_c$。并在 $Y_c$ 的指导下将其分配为正（来自非显著区域）或负（来自背景区域），全局和局部类令牌分别通过映射 $G$ 和 $F$，且皆由线性层和 $L2$ 归一化层组成，假设 $p$ 表示映射全局类令牌，$Q^+ / Q^-$ 表示从非显著区域或者背景区域裁剪的映射局部的类令牌，CTC 的目标是最小化/最大化 $p$ 与 $Q^+ / Q^-$ 差异，使用 InfoNCE 损失[20-23]作为目标函数，定义公式如下：

$$L_{\mathrm{CTC}} = \frac{1}{N^+}\sum_{q^+}\log\frac{e^{\left(\frac{p^{\mathrm{T}}q^+}{\tau}\right)}}{e^{\left(\frac{p^{\mathrm{T}}q^+}{\tau}\right)} + \sum_{q^-} e^{\left(\frac{p^{\mathrm{T}}q^-}{\tau}\right)} + \epsilon} \tag{2}$$

其中，$q^+ \in Q^+$；$q^- \in Q^-$；$N^+$ 计算 $q^+$ 的个数；$\tau$ 表示温度因子；$\epsilon$ 表示稳定性的一个小正值。CTC 目的是强制局部视图表示与全局视图表示保持一致。因此，停止投影头 $G$ 的梯度。为了更新 $G$ 采用指数移动平均 $\theta^g \leftarrow \rho\theta^g + (1-\rho)\theta^l$，其中 $\rho$ 是动量因子，这里设置为 0.9，$\theta^g$ 和 $\theta^l$ 分别是来自 $G$ 和 $F$ 的参数。

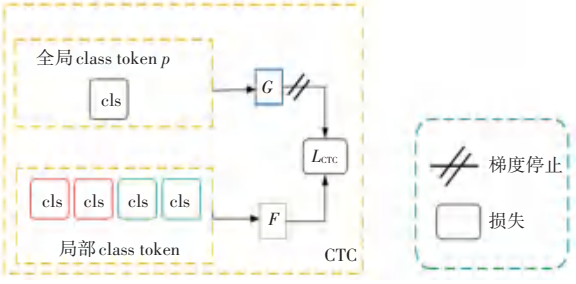


**图 3　类令牌对比模块**

**Fig. 3　Class token contrast module**

## 2.3　SAM 下的伪标签生成

SAM 在做弱监督语义分割时对于能见度差的隐藏目标，例如从稀疏标注的蓝膜数据集中学习分割模型 $S = \{X_i, Y_i\}_{i=1}^{S}$，然后在测试数据集上测试模型 $T = \{T_i\}_{i=1}^{T}$，这里的 $X_i$ 和 $T_i$ 分别表示训练图像和测试图像；$Y_i$ 表示稀疏注释，可以是几个点或作为前景或背景注释的涂鸦。由于稀疏注释 $Y_i$ 无法提供足够的监督去训练出准确密集预测的模型，为了从稀疏注释中生成高质量的密集掩码，提出了 2 种策略改善 SAM 生成的掩码质量，即图像增强结果融合、像素加权，SAM 生成的融合图如图 4 所示。2 种策略如下：

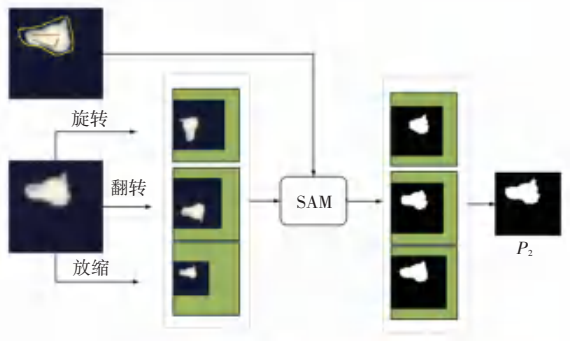


**图 4　SAM 引导融合过程**

**Fig. 4　SAM-guided fusion process**

（1）数据增强结果融合。给定一个蓝膜图像 $(X_i, Y_i) \in S$，首先通过应用随机增广生成 $K$ 张增强图 $\{X_i^k\}_{k=1}^{K}$，从图像翻转、旋转中随机采样（90°，180°，270°，360°）和放缩（0.5，1.0，2.0），训练使用稀疏注释 $Y_i$ 作为提示，将 $\{X_i^k\}_{k=1}^{K}$ 发送给 SAM，生成分割蒙版 $\{M_i^k\}_{k=1}^{K}$ 这里 $M_i^k = \mathrm{SAM}(X_i^k, Y_i)$，$M_i^k$ 与输入图像 $X_i^k$ 应具有相同的形状，当形状上有不同时 $M_i$ 则进行图像逆变换，以确保所有蒙版大小与原始图像相同。由于给 SAM 输入不同的提示会得到不同的分割结果，不同的增强图像得到了不同的分割蒙版。虽然这些掩模的形状变化很大，但却在某些区域重叠，无论图像变换如何，SAM 都能可靠地预测正确的前景区域。另外，这些蒙版是相互补充的，这样一个蒙版遗漏的前景区域可以在其他蒙版中找到。融合不同增强图像的分割掩码，则可以表示为：

$$\tilde{M}_i = \frac{1}{K}\sum_{k=1}^{K} M_i^k \tag{3}$$

其中，$\tilde{M}_i$ 表示融合后的的伪掩码。

（2）像素权重。不同像素的预测结果可靠性是不同的，使用熵来衡量预测结果的可靠性，计算每个像素的熵，得到一个熵图，具体公式如下：

$$\tilde{E}_i = -\tilde{M}_i \log \tilde{M}_i - (1-\tilde{M}_i)\log(1-\tilde{M}_i) \tag{4}$$

由于熵图是由融合掩模计算的，测量了所有增强图像中每个像素的预测不确定性，因为只有当一个像素被一致地从所有增强图像中预测时才会具有低熵。使用这个熵图来衡量融合掩模 $M_i^k$，并为那些预测可靠的像素分配更高的权重。

## 2.4　伪标签特征融合

$P1$ 和 $P2$ 两张伪标签表示同一图片的分割目标，SAM 生成的伪标签侧重于特征突出目标区域，而 ViT 生成的伪标签侧重于整个不同于背景的图像

区域,通过卷积网络将由 SAM 生成的伪标签和由 ViT 网络生成的伪标签进行融合得到更接近于真实标签的伪标签。融合方式如图 5 所示。

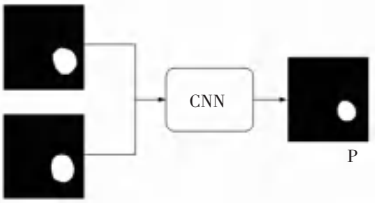


**图 5 伪标签融合**

**Fig. 5 Pseudo-label fusion**

## 3 实验与分析

### 3.1 实验设置及评估标准说明

为了验证本文所提模型的性能,使用蓝膜数据集。其中,缺陷种类有划痕、泡棉、压痕、异物各 300 张、共 1 200 张图片。在训练阶段,所有对比实验和本文提出的模型均采用 adamW 优化器,学习率为 0.000 06,动量和权重衰减分别设置成 0.9 和 0.001,默认批量大小为 8。

(1)硬件和软件环境:所有实验均在单个 Nvidia RTX3090 GPU 进行,使用 Anaconda 深度学习平台,Pytorch 深度学习框架。

(2)评价指标:在评估模型性能时,使用了 2 个指标:ACC 和 mIoU。

① 准确率(ACC)。是衡量模型在所有像素上的正确分类数与总像素数之比。在语义分割任务中,准确率计算预测像素与真实像素之间的匹配情况。具体计算公式如下:

$$ACC = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

其中,TP 表示真正例的数量;TN 表示真负例的数量;FP 表示假正例的数量;FN 表示假负例的数量。

② mIoU。是常用的分割任务评价指标,用于衡量模型对每个类别的分割结果的准确性。对于每个类别,计算预测像素和真实像素的交集面积和并集面积之比,然后对所有类别取平均得到 mIoU。具体计算公式如下:

对于每个类别 $i$, 计算其 IoU(Intersection over Union),定义公式为:

$$IoU_i = TP_i/(TP_i + FP_i + FN_i) \quad (6)$$

其中, $TP_i$ 表示类别 $i$ 中真正例(True Positive)的数量; $FP_i$ 表示类别 $i$ 中假正例(False Positive)的数量; $FN_i$ 表示类别 $i$ 中假负例(False Negative)的数量。

取所有类别的 IoU 的平均值,得到 mIoU 的计算公式如下:

$$mIoU = (IoU\_1 + IoU\_2 + \cdots + IoU\_C)/C \quad (7)$$

其中, $C$ 表示类别数量。

### 3.2 对比实验

为了评估模型结构的有效性,本文对比了被广泛应用的弱监督语义分割网络。不同网络对比实验结果见表 1。

**表 1 不同网络对比实验**

**Table 1 Comparative experiments of different networks**

| 模型 | ACC ↑ | mIoU ↑ | Params/M | GFLOPs |
|---|---|---|---|---|
| RIB | 84.49 | 79.16 | 68.59 | 210 |
| ReCAM | 85.12 | 82.49 | 50.42 | 198 |
| SEAM | 85.63 | 80.23 | 60.28 | 177 |
| AFA | 85.13 | 81.29 | 87.69 | 363 |
| SLRNet | 87.11 | 82.29 | 90.32 | 444 |
| 本文 | 89.51 | 85.65 | 58.94 | 236 |

表 1 的实验数据证明,本研究提出的方法在精确度和平均分割率评估指标上表现优于其他方法。

### 3.3 消融实验

为了探索不同因素对模型性能的影响,进行了消融研究。实验如下。

(1)针对 SAM 引导提供的数据增强效果。实验结果见表 2。

**表 2 由 SAM 生成伪标签的方式**

**Table 2 Generation of pseudo-labels by SAM**

| SAM 引导方式 | ACC ↑ | mIoU ↑ |
|---|---|---|
| 无 SAM | 86.96 | 83.68 |
| SAM 图像翻转 | 88.10 | 85.01 |
| SAM 图像伸缩 | 88.23 | 85.12 |
| SAM 图像融合 | 89.51 | 85.65 |

在表 2 中实验数据表明通过融合数据增强的 SAM 伪掩码结果可以有效地改善伪标签的质量。

(2)针对各个模块的有效性结果验证。结果见表 3。

**表 3 CTC 模块和 SAM 模块的影响**

**Table 3 Impact of CTC module and SAM module**

| M_total | ACC ↑ | mIoU ↑ |
|---|---|---|
| M1 | 85.96 | 82.29 |
| M1+M2 | 86.18 | 83.29 |
| M1+M3 | 88.11 | 84.29 |
| M1+M2+M3 | 89.51 | 85.65 |

在表 3 中,M1、M2、M3 分别表示提出的模型中的 ViT 基础模块、类令牌对比模块以及多增强 SAM

引导模块。

### 3.4　结果对比展示

对于几种缺陷分割效果的对比如图 6 所示。图 6 中，$P_1$ 表示 ViT 过程产生的伪标签，$P_2$ 表示 SAM 产生的伪标签，$P$ 表示两者融合后的伪标签。实验结果表明本文方法对大多数缺陷的分割伪标签都有所改进，只有很少的隐藏缺陷在伪标签融合后效果不如 $P_1$，但是总体上伪标签 $P$ 与真实的标签 GT 是最相近的，说明了本文方法的有效性。

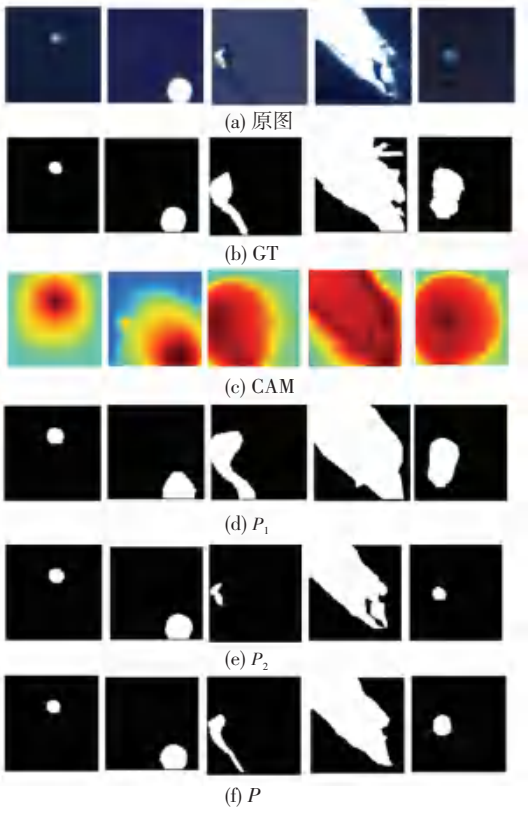


**图 6　分割效果对比图**

**Fig. 6　Comparison of segmentation results**

## 4　结束语

本文提出一种用于电芯蓝膜表面缺陷检测的伪标签生成算法。该算法采用 ViT 特征提取网络添加类令牌对比模块得到伪标签 $P1$；采用 SAM 模型，通过数据增强结果融合和像素加权来生成伪标签 $P2$，将 2 种伪标签通过 CNN 网络融合成精细化的伪标签 $P$，在蓝膜表面缺陷数据集上进行了测试。该算法仅使用图像级别的标签就可以预测出缺陷的种类和位置，其中类令牌对比模块提升了网络识别表面缺陷完整度，SAM 引导生成模块提升了网络分割真实缺陷的能力，使得模型能够精准地定位到图像中缺陷区域。在工业缺陷检测要求不高的情况下，相比于全监督算法，利用本文算法可以实现快速判别并定位出的缺陷所在的大致区域，省去了繁琐的人工标注。但是由于弱监督算法使用的是信息量更少的类别标签，本文分割算法分割精度上会稍落后于全监督的分割算法，因此，弱监督的缺陷检测算法还存在很大的提升空间，值得深入地展开进一步研究。

## 参考文献

[1] 唐东昂. 彩色图像下的圆柱形覆膜锂电池圆周面缺陷检测方法研究[D]. 沈阳：沈阳工业大学，2021.

[2] 胡维鑫，尹佳，田辉. 基于双目视觉的航空复杂结构件机器人制孔研究[J]. 工具技术，2023，57(12)：101-105.

[3] 陶显，侯伟，徐德. 基于深度学习的表面缺陷检测方法综述[J]. 自动化学报，2021，47(5)：1017-1034.

[4] 许斌，肖宇萌，武玉洁，等. 基于深度学习的陶瓷薄片表面划痕检测方法[J]. 工具技术，2022，56(8)：142-146.

[5] KOLESNIKOV A，LAMPERT C H. Seed，expand and constrain：Three principles for weakly-supervised image segmentation[C]// Proceedings of the European Conference on Computer Vision. Cham：Springer，2016：695-711.

[6] HUANG Zilong，WANG Xinggang，WANG Jiasi，et al. Weakly-supervised semantic segmentation network with deep seeded region growing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE ，2018：7014-7023.

[7] FENG Jiapei，WANG Xinggang，LIU Wenyu. Deep graph cut network for weakly - supervised semantic segmentation [J]. Science China Information Sciences，2021，64(3)：130105.

[8] AHN J，KWAK S. Learning pixel-level semantic affinity with image - level supervision for weakly supervised semantic segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2018：4981-4990.

[9] AHN J，CHO S，KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2019：2209-2218.

[10] WANG Yude，ZHANG Jie，KAN Meina，et al. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2020：12275-12284.

[11] SHEN Tong，LIN Guosheng，LIU Lingqiao，et al. Weakly supervised semantic segmentation based on co-segmentation[C]// Proceedings of the 28th British Machine Vision Conference (BMVC). London：BMVA Press，2017：17.

[12] WEI Yunchao，LIANG Xiaodan，CHEN Yunpeng，et al. STC：A simple to complex framework for weakly supervised semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，39(11)：2314-2320.

[13] CHEN L C，PAPANDREOU G，KOKKINOS I，et al. Deeplab：Semantic image segmentation with deep convolutional nets，atrous convolution，and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(4)：834-848.

[14] SONG Chunfeng, HUANG Yan, OUYANG Wanli, et al. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 3136-3145.

[15] ROSSETTI S, ZAPPIA D, SANZARI M, et al. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation [C]//Proceedings of the European Conference on Computer Vision. Cham: Springer, 2022: 446-463.

[16] CHEN Qi, YANG Lingxiao, LAI Jianhuang, et al. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 4288-4298.

[17] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[J]. arXiv preprint arXiv, 2012.12877, 2020.

[18] LI Ruiwen, MAI Zheda, TRABELSI C, et al. Transcam: Transformer attention-based CAM refinement for weakly supervised semantic segmentation[J]. arXiv preprint arXiv, 2203.07239, 2022.

[19] KIRILLOV A, MINTUN E, RAVI N, et al. Segment anything [J]. arXiv preprint arXiv, 2304.02643, 2023.

[20] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. TPAMI, 2017, 40(4): 834-848.

[21] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[J]. arXiv preprint arXiv, 1412.7062, 2015.

[22] LEE J, OH S J, YUN S, et al. Weakly supervised semantic segmentation using out-of-distribution data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16897-16906.

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