文章编号　1004-924X(2026)01-0150-17

时空对比学习驱动的弱监督图像语义分割网络

梁　臻[1,2]，胡燕祝[1,2*]，杨　洋[1,2]

(1. 北京邮电大学 智能工程与自动化学院，北京 100876；
2. 北京邮电大学 物联网监测预警应急管理部重点实验室，北京 100876)

摘要：现有基于视觉变换器(Vision Transformer，ViT)的图像级弱监督语义分割方法主要依赖自注意力机制提取有限的语义信息，未能充分利用多维特征关系，导致目标区域的识别较为粗略。为此，提出一种时空对比学习驱动的弱监督图像语义分割网络(Spatio-temporal Contrastive Learning，STCL)，旨在通过时间、空间角度挖掘监督信息，以提高分割精度。通过ViT的令牌机制，引入了空间特征对比学习模块，结合补丁级令牌和类级令牌对比策略，深入探索图像空间中隐含的语义特征关系；设计了时间上下文对比学习模块，通过构建记忆库，利用历史图像分割中的先验知识来指导当前语义分割任务，并建立了记忆库更新策略和自适应记忆对比度损失，进一步提升了模型对细节区域的辨识能力。实验结果表明，在PASCAL VOC和MS COCO数据集上的平均交并比可以分别达到72.7%以及43.6%，证明了所提出方法的优越性。

关　键　词：计算机视觉；语义分割；弱监督学习；类激活图；视觉变换器；对比学习

中图分类号：TP391　　**文献标识码**：A

doi：10.37188/OPE.20263401.0150　　**CSTR**：32169.14.OPE.20263401.0150

Weakly supervised image semantic segmentation network driven by spatio-temporal contrastive learning

LIANG Zhen[1,2], HU Yanzhu[1,2*], YANG Yang[1,2]

(1. *School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *Key Laboratory of IoT Monitoring and Early Warning, Ministry of Emergency Management, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)
* *Corresponding author, E-mail: bupt_automation_safety_yzhu@bupt.edu.cn*

Abstract: Existing image-level weakly supervised semantic segmentation methods based on Vision Transformer (ViT) primarily rely on self-attention to extract limited semantic information and often fail to fully exploit multi-dimensional feature relationships, resulting in coarse target region identification. To address this limitation, a Spatio-temporal Contrastive Learning network (STCL) is proposed to improve segmentation accuracy by mining supervisory signals from both spatial and temporal perspectives. Specifically, a spatial feature contrastive learning module is introduced based on ViT token representations, integrating

收稿日期：2025-09-10；**修订日期**：2025-09-30.

基金项目：国家自然科学基金资助项目(No. 52478123)；国家重点研发计划资助项目(No. 2020YFC1511700)；北京邮电大学博士生创新基金资助项目(No. CX20242080)

patch-level and class-level token contrastive strategies to capture implicit semantic relationships in image space. In addition, a temporal context contrastive learning module is developed, in which a memory bank is leveraged to incorporate prior knowledge from historical images to guide current segmentation, together with a memory bank update strategy and an adaptive memory contrastive loss to enhance discrimination of fine-grained regions. STCL achieves a mean Intersection over Union (mIoU) of 72.7% on PASCAL VOC and 43.6% on MS COCO, demonstrating superior performance.

Key words: computer vision; semantic segmentation; weakly supervised learning; class activation map; vision transformer; contrastive learning

1　引　言

语义分割主要研究计算机视觉中的像素级目标识别。由于语义分割生成的预测结果包含丰富的场景理解信息，该技术在自动驾驶及医学图像分析等领域[1-2]得到了广泛应用。随着深度学习技术的兴起，全监督语义分割取得了显著的进展[3-4]。然而，这类任务通常依赖于精细的像素级注释，既耗时又昂贵[5]。在这种情况下，弱监督语义分割越来越受到关注，它利用粗糙的标注进行分割，如边界框[6-7]、点[8]、涂鸦[9-10]和图像级标签[11]。其中，图像级标签是最具成本效益的标注形式，也是弱监督语义分割场景中最具挑战性的任务。

目前，大多数解决该任务的方法是利用神经网络生成类激活图（Class Activation Map，CAM）来定位图像中的目标，基于CAM获得伪标签来训练语义分割网络。CAM通过对分类网络中最后一个卷积层的特征图进行加权求和，生成类别激活图，从而实现弱监督下的物体定位和视觉解释[12]。原始CAM只激活最具分辨性的区域，导致物体边界模糊和不完整。因此，如何获得高质量的CAM已成为弱监督语义分割的核心问题。随着深度学习的发展，很多图像级弱监督语义分割方法使用CNN作为骨干网络来优化CAM[13]。但是由于卷积运算无法进行全局特征建模，该策略只能实现有限的改进。近年来，随着视觉变换器（Vision Transformer，ViT）在视觉识别任务中不断取得突破，出现了以ViT为基础的弱监督语义分割的解决方案。ViT通过自注意力机制捕获全局语义依赖，解决了CNN的局限性，展示了在弱监督语义分割任务中的巨大潜力。基于此，研究人员试图优化ViT的结构[14-15]，探索令牌的潜在语义信息[16-18]，并通过挖掘图像内部监督提高分割精度[19-20]。这些方法大多侧重于通过挖掘图像空间中固有的语义特征关系来获得有监督的信息。由于图像级标签只能提供粗粒度的信息，单纯从特征关系中获得的监督信息是有限的，制约了语义分割的性能。因此，有必要从先验知识中提取更丰富全面的监督信息，以指导模型进行精确的弱监督语义分割。

针对这一挑战，本文提出了一种时空对比学习驱动的弱监督图像语义分割网络（Spatio-temporal Contrastive Learning，STCL），利用时间和空间维度的细粒度先验来指导图像精确分割。从ViT的令牌机制中挖掘足够的空间监督信息，从历史图像的分割中吸收目标的类别特征知识，以此来提高图像弱监督语义分割的性能。首先，引入空间特征对比学习模块来探索CAM优化中隐含的令牌特征语义信息，通过构建辅助分类器，利用原始和辅助补丁级令牌之间的差异激活CAM区域。其次，设计了全局空间和局部空间类级令牌之间的一致性策略来进一步改进CAM。然后，构建了一个时间上下文对比学习模块，以充分捕捉历史上下文语义相关性，通过设计记忆原型提取类别特征，构建初始记忆库存储历史知识，并采用记忆库更新策略在线更新历史先验信息，指导CAM持续优化。最后，利用改进后的CAM生成伪标签进行图像分割，并训练ViT语义分割网络，生成最终的分割结果。实验结果表明，所提出的方法优化了CAM，与现有的弱监督语义分割方法相比，分割性能更佳。

2　相关工作进展

弱监督语义分割是一项具有挑战性的任务，已成为近年来的研究热点，主要分为图像级标签

的弱监督语义分割方法和基于ViT的弱监督语义分割方法。

2.1　图像级弱监督语义分割

目前,根据CAM优化阶段的不同,图像级弱监督语义分割可分为多阶段方法和单阶段方法。

2.1.1　多阶段方法

多阶段弱监督语义分割是指先通过训练分类网络获得CAM,再通过CAM生成的伪标签训练语义分割网络。CAM的优化是该任务的核心,目前大多数图像级弱监督语义分割方法都是多阶段方法。为了生成高质量的CAM,研究人员提出了许多相关方法。部分方法通过优化网络结构来提高CAM的定位精度[21-23],文献[23]通过减少信息瓶颈改善了弱监督语义分割中的目标定位。一些方法通过利用模型中的特征关系来提高伪标签的质量[16-17],文献[17]学习了类令牌和补丁令牌之间的交互,以捕获每个类的独特特征。此外,大量方法通过额外的策略来挖掘图像中的语义信息[12,24-27],文献[12]通过像素到原型的对比学习缩小了分类和分割之间的监督差距。还有一些方法通过亲和学习来预测语义信息[28-29],使用外部数据来增强目标信息[30-32],或致力于解决CAM误激活问题[33-34]。多阶段方法允许在每个阶段引入不同的技术,为调整训练过程提供了灵活性。然而,它们的处理过程更复杂,需要更多的时间和计算资源。

2.1.2　单阶段方法

与多阶段方法不同,单阶段方法旨在直接使用图像级标签作为监督来训练端到端分割网络。文献[35]通过生成可靠的像素级标注实现了高效的分割。文献[36]和文献[37]使用自监督学习来补充弱监督语义分割的语义信息。这些方法一般采用CNN作为骨干网络,但由于卷积固有的局限性,它们只能获得不完全的对象激活效果[38]。随着Transformer的发展,基于ViT的单阶段弱监督语义分割方法应运而生。一些方法利用ViT挖掘令牌语义信息[18,39]。文献[28]和[40]引入了亲和学习来生成更完整的伪标签。此外,还有方法结合蒸馏学习[41]来优化CAM。这些方法简化了训练过程,通常需要较少的计算资源。然而,现有方法通常取得较为有限的性能,这主要源于它们往往依赖简单的语义先验来指导图像分割过程。为了解决这个问题,本文提出了STCL网络充分探索时间和空间维度中隐含的语义信息,以此来进行高质量的分割。

2.2　基于ViT的弱监督语义分割方法

借助ViT强大的注意力机制,许多基于ViT的弱监督语义分割方法被提出,赋予了模型强大的目标定位能力,并取得了令人印象深刻的结果。根据监督策略的不同,现有的基于ViT的弱监督语义分割方法可分为三种类型:令牌级监督、模型级监督和图像级监督。具体介绍如下:

(1)令牌级监督:这类解决方案利用ViT中的内在补丁和类令牌表达为弱监督语义分割中的细粒度语义监督学习建模特征关系,并已成为最近研究的主流策略。文献[39]通过利用全局最大池化和ViT的补丁特征编码能力构建了一个端到端的补丁-类令牌映射关系。文献[16]使用多个类令牌来学习类令牌和补丁令牌之间的交互,以捕捉不同类别的特征。文献[18]引入了补丁令牌对比和类令牌对比模块,以解决ViT中的过度平滑问题。

(2)模型级监督:这种方法通过引入额外的模块或多分支架构来探索模型中的监督信息,以指导特征学习和伪标签细化。文献[41]对注意力机制生成的特征图之间的对应关系进行了自监督蒸馏,以提高语义信息的完整性。文献[40]对ViT模型中的自注意力机制进行建模和监督,以挖掘隐含的语义关系。

(3)图像级监督:这种方法的核心思想是在图像内部挖掘监督知识,以提高弱监督语义分割的性能。文献[19]引入了亲和一致性,探索图像内区域的成对关系,以确保每个视图内不同区域之间的相关性保持稳定。文献[33]结合了一种对抗学习机制,通过探索图像中不同区域之间的信息依赖关系来捕获细粒度的监督信号。

基于上述分析,目前基于ViT的弱监督语义分割方法主要侧重于利用ViT的独特架构通过单源特征学习挖掘监督知识。然而,由于图像级标签只提供粗粒度信息,仅依赖从特征相关性中获得的这种单源监督具备固有的局限性,这无疑严重限制了语义分割的性能。因此,为了解决这一差距,本文提出了一种时空对比学习驱动的弱监督图像语义分割网络,从时间维度和空间维度挖掘丰富的监

督信息，以获得准确的弱监督语义分割结果。

3　原　理

本文提出的一种时空对比学习驱动的弱监督图像语义分割网络，可从时间角度和空间角度提取丰富的监督信息，以实现准确的弱监督语义分割。如图1所示，该方法由两个核心模块组成：空间特征对比学习模块和时间上下文对比学习模块。

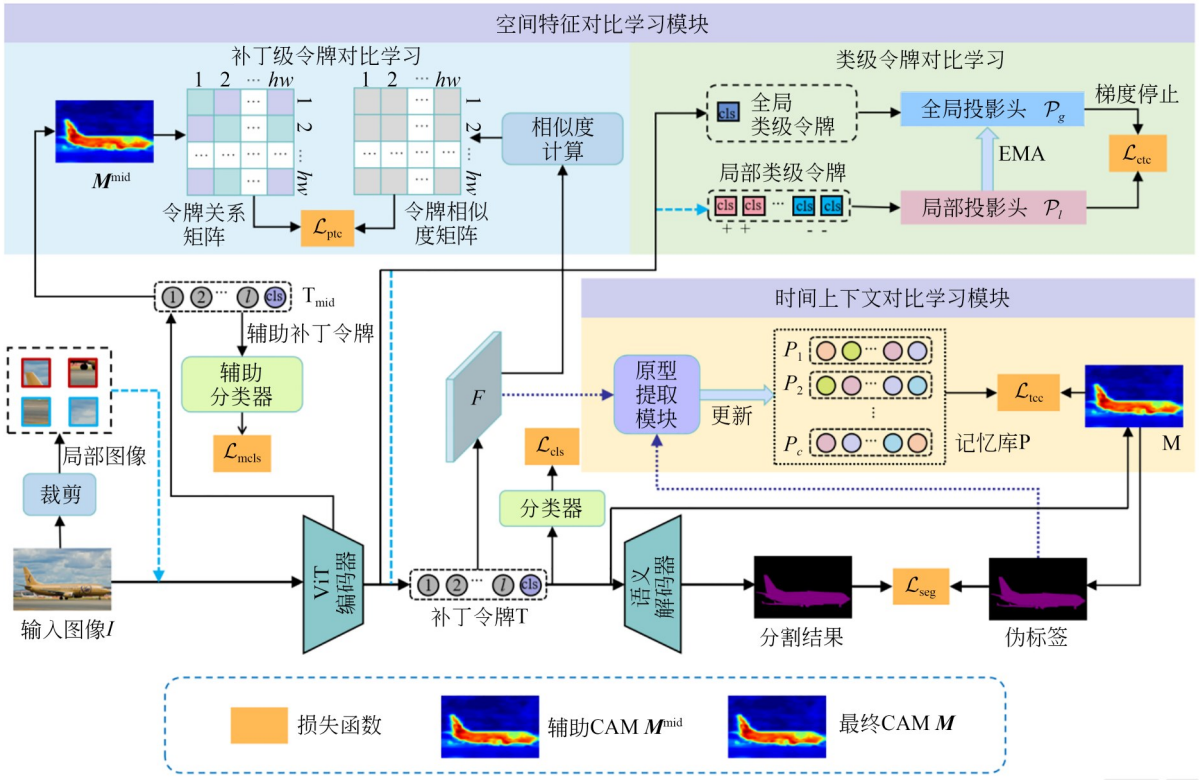

图1　时空对比学习驱动的弱监督图像语义分割网络框架

Fig. 1　Framework of weakly supervised semantic segmentation network driven by spatio-temporal contrastive learning

3.1　准备工作

CAM由于有效性和简单性被广泛应用于像素级弱监督语义分割中，本文利用CAM生成图像的初始伪标签。给定一张图像$I\in\mathbf{R}^{3\times H\times W}$，视觉变换器将输入图像划分成固定大小的补丁令牌，若将初始补丁令牌的尺寸设为S，则补丁令牌的数量为$(H/S)\times(W/S)$。将这些令牌序列与一个可学习的类令牌结合起来，经过视觉变换器的编码器，可以得到最终的令牌序列，表示为$T_{out}\in\mathbf{R}^{D\cdot((H/S)\cdot(W/S)+1)}$，其中，$D$代表嵌入维度。然后，经过卷积操作，得到特征图$F\in\mathbf{R}^{C\cdot(H/S)\cdot(W/S)}$。其中，$C$代表输出通道，即类别数目。通过对特征图进行加权求和，应用ReLU函数和最大归一化操作后，可以得到对于c类别的初始CAM，如式(1)所示。其中，F代表特征图，W代表权重。

$$\mathrm{CAM}_c=\frac{\mathrm{ReLU}(M_c)}{\max(\mathrm{ReLU}(M_c))},M_c=\sum_i^c W_i F_i. \quad (1)$$

本文从空间角度和时间角度，利用ViT中的令牌机制和记忆库历史原型匹配模块，优化CAM的生成，然后利用CAM生成伪标签，将改进后的伪标签将用于监督分割解码器，进而优化图像分割结果。

3.2　空间特征对比学习模块

利用ViT框架的令牌机制，通过空间特征对比学习模块，探索图像空间中丰富的语义特征相关性来优化CAM，该模块分为补丁级令牌对比学习和类级令牌对比学习。

3.2.1　补丁级令牌对比学习

补丁级令牌对比模块利用不同补丁令牌之间的关系，为弱监督语义分割提供监督信息。给

定输入图像I,通过ViT编码器构造补丁令牌,其中编码器由多个Transformer块组成,每个Transformer块输出一系列补丁令牌。因此,在一个模型推理中可以得到ViT块$T=\left[T_i\in\mathbf{R}^{n\times d}\right]_{i=1}^{N}$的不同层的令牌,其中$N$为Transformer块的数目,$n$和$d$分别为令牌个数和特征维度。通常,CAM是由最终的补丁令牌生成的。相关研究表明,前期层中的语义信息不够丰富,后期层中的表征特征往往是一致的,这无疑限制了现有方法的性能。在此基础上,设计了补丁级令牌对比策略,从中间层引入语义特征,来增强最终补丁令牌的目标区分能力,以改善弱监督语义分割。为了实现这一目标,在中间层添加一个辅助分类器来增加辅助CAM,以聚集语义特征,并利用获得的中间层特征来优化最终的补丁令牌。

具体来说,将中间层的补丁令牌T_{mid}进行整合,通过全局最大池化操作生成辅助CAM $\boldsymbol{M}^{\text{mid}}$。然后设置$\alpha,\beta$两个阈值,将$\boldsymbol{M}^{\text{mid}}$划分为背景、前景和不确定区域,进而得到辅助CAM的伪标签$\boldsymbol{L}_{\text{mid}}\in\mathbf{R}^{h\times w}$。在此基础上,设计损失函数,根据$\boldsymbol{L}_{\text{mid}}\in\mathbf{R}^{h\times w}$,增大最终补丁令牌对不同目标的特征差异区分能力,并增强最终补丁令牌对相同对象的特征表达一致性。首先,从最终的补丁令牌T中构建图像的二维语义特征图F,然后通过计算F中不同令牌之间的绝对余弦相似度获得令牌相似度矩阵,之后利用伪令牌标签$\boldsymbol{L}_{\text{mid}}\in\mathbf{R}^{h\times w}$建立令牌关系矩阵,以此来优化令牌相似度。如果$T_{\text{mid}}$中的两个令牌共享相同的语义标签,则将它们标记为正对,用于构建正掩码$\mathcal{M}^{+}\in\mathbf{R}^{hw\times hw}$。相反,如果$T_{\text{mid}}$中的两个令牌属于不同的语义标签,则将它们标记为负对,用于构建负掩码$\mathcal{M}^{-}\in\mathbf{R}^{hw\times hw}$。为了确保可靠性,重点关注前景和背景,忽略不确定区域的令牌。基于这种设置,损失函数被设计成以下形式,以最大化正对补丁令牌的相似性,最小化负对补丁令牌的相似性。

$$\mathcal{L}_{\text{ptc}}=\frac{1}{2}\left(1-\frac{\sum_{i,j}\mathcal{M}^{+}\circ\text{ASim}\left(\boldsymbol{F}_i,\boldsymbol{F}_j\right)}{N^{+}}\right)+\frac{1}{2}\left(\frac{\sum_{i,j}\mathcal{M}^{-}\circ\text{ASim}\left(\boldsymbol{F}_i,\boldsymbol{F}_j\right)}{N^{-}}\right),\tag{2}$$

其中:N^{+}和N^{-}分别表示正对和负对的个数,$\circ$表示Hadamard乘积运算,ASim表示绝对余弦相似度的计算。通过促进正令牌对趋于一致,负令牌对更加不同,中间层令牌能为最终层令牌提供有效监督,进而细化CAM。

3.2.2 类级令牌对比学习

CAM的一个关键限制是全局图像空间与局部图像空间的语义表达的不一致。为了缓解这个问题,利用ViT中类令牌可以聚合高级语义的特性,设计了一个类级令牌对比机制去增强局部区域和全局对象之间的表示一致性,促进CAM激活更多的对象区域。

基于上述辅助CAM生成的伪标签$\boldsymbol{L}_{\text{mid}}$,首先,从原始图像中将不确定和背景区域进行裁剪,获得一系列局部图像。然后,将这些局部图像和原始图像输入到ViT编码器中,以获得局部类令牌和全局类令牌。类令牌可以有效地表示对象的语义信息[38,42]。受此启发,本文致力于最小化全局类令牌与不确定的局部类令牌之间的差异,并最大化全局类令牌与背景局部类令牌之间的差异,以有效地激活CAM的对象区域,促进整个对象区域的更一致表示。具体来说,采用全局投影头$\mathcal{P}_{\text{g}}$和局部投影头$\mathcal{P}_{\text{l}}$来投影全局和局部令牌,这些投影头均由线性层和L2范数归一化组成。假设t^{l}代表投影的全局类令牌,t^{+}和t^{-}分别代表投影的背景和不确定区域的令牌,本文使用InfoNCE损失[43]设计以下损失函数$\mathcal{L}_{\text{ctc}}$来指导模型的优化,表示为:

$$\mathcal{L}_{\text{ctc}}=\frac{1}{N^{+}}\sum_{t^{+}}\log\frac{\exp\left(\frac{t^{l}t^{+}}{\tau}\right)}{\exp\left(\frac{t^{l}t^{+}}{\tau}\right)+\sum_{t^{-}}\exp\left(\frac{t^{l}t^{-}}{\tau}\right)+\varepsilon},\tag{3}$$

其中:N^{+}表示t^{+}的个数,τ是温度因子,ε是一个非常小的正数,以防止分母为零。特别地,为了确保局部令牌与全局令牌的对齐,停止全局投影头$\mathcal{P}_{\text{g}}$的梯度,并使用指数移动平均(Exponential Moving Average, EMA)方法更新其参数,从而适应数据变化并增强训练过程的稳定性。

3.3 时间上下文对比学习模块

现有的基于ViT的弱监督语义分割方案一

般只利用单个图像中有限的上下文线索，没有从时间角度吸收数据集中其他图像的分割结果，无法全面地挖掘语义知识。为了使历史图像为后续的分割提供参考和指导，本文设计了一种时间上下文对比学习模块，通过构建记忆库，用原型保存每个目标类的语义特征作为监督信息，指导后续图像吸收先验知识，从而优化分割效果。为了进一步展示记忆库的更新过程，图 2 展示了详细的时间上下文对比学习模块流程图。

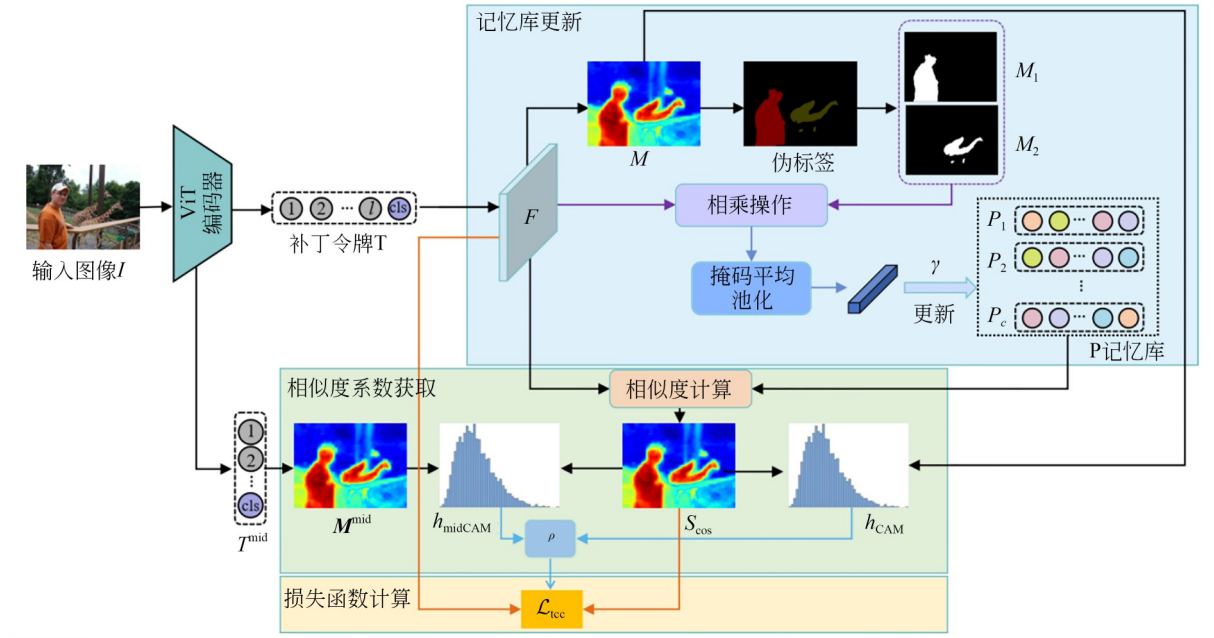

图 2　时间上下文对比学习模块流程

Fig. 2　Flowchart of temporal context contrastive learning module

具体来说，构建记忆库 $\boldsymbol{P}=\{P_1,P_2,\cdots,P_C\}$ 来保存对象类别的原型信息，它由 C 个原型字典组成，其中 P_l 表示类别 l 的原型，C 为类别总数。通过将记忆库中的类别特征与图像特征进行匹配，可以对齐时间上下文特征的一致性，从而进一步细化 CAM。当获得图像的特征图 $\boldsymbol{F}$ 和 CAM $\boldsymbol{M}$ 时，根据 3.2 节设置的阈值将 CAM 划分为背景、前景和不确定区域，并使用 CAM 生成的伪标签获取图像中的类别信息。利用伪标签，为每个目标类别区域构建一个二元掩膜。对于类别 l，掩膜可以表示为：

$$\mathcal{M}_{i,j}^{l}=\begin{cases}1, & \text{if}\quad \text{label}_{i,j}==l\\ 0, & \text{otherwise.}\end{cases}\tag{4}$$

如果伪标签中位置 (i,j) 的像素属于类别 l，则对应的掩膜位置设为 1，其余设为 0。利用该掩膜将每个目标类元素对应的特征图中的语义特征提取出来，并添加到原型列表中的对应元素，作为初始原型信息。从特征图中提取图像中类别 l 的原型 P_l，表示为：

$$P_l=\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}F_{i,j}\cdot\mathcal{M}_{i,j}^{l}}{\sum_{i=1}^{H}\sum_{j=1}^{W}\mathcal{M}_{i,j}^{l}+\varepsilon},\tag{5}$$

其中：$F_{i,j}$ 表示特征图中 (i,j) 处的值，$\mathcal{M}_{i,j}^{l}$ 为类别 l 在位置 (i,j) 处的掩码值，ε 为防止分母为零的小正数。式(5)通过将每个类的掩膜与特征图相乘，将每个类对应的特征提取出来，然后进行掩码平均池化操作(mask Average Pooling，mAP)，将特征转为一维向量，此时获取的一维掩码特征便是对应类的原型，最后将其添加到原型列表中，组成记忆库。通过在线迭代不断获取不同类别的原型，逐步构建完整的记忆库。

为了在有限的内存空间内从大量样本中获

取准确的原型类别信息，本文设计了原型更新机制，更新记忆库中每个类别的原型。记忆库将在反向传播阶段不断更新，吸收有用的语义信息，增强对目标的语义理解。以类别 l 为例，最新的原型 P_l' 可以通过以下过程进行更新：

$$P_l' \leftarrow \gamma P_l^b + (1-\gamma) P_l, \tag{6}$$

其中：P_l^b 为更新前的原型，P_l' 为更新后的原型，P_l 代表当前图像中目标类别的特征信息，γ 为动量因子，用作记忆更新策略的权重，以保证更新的稳定性。该更新策略并非在每个模型推理中均进行更新，而是根据特定的频率进行处理。

基于上述设置，设计了一种时间上下文对比损失，通过对比输入图像和历史目标特征之间的差异来提取监督信息。具体来说，首先计算输入图像特征与记忆库中每个原型之间的余弦相似度，生成特征相似度图 $\boldsymbol{S}_{\cos}$ 来初步定位对象，如下：

$$\begin{aligned}\boldsymbol{S}_{\cos} &= \mathrm{Sim}_{\cos}(P, F) = \frac{\boldsymbol{P} \cdot \boldsymbol{F}}{\| \boldsymbol{P} \| \cdot \| \boldsymbol{F} \|}, \\ \boldsymbol{P} &= \{P_1, P_2, \cdots, P_c\}.\end{aligned} \tag{7}$$

由于原型表示类别信息，类似于 CAM，相似度图也可以反映像素属于对象类别的概率。因此，相似度图能够发挥类似于辅助 CAM 的作用，进而可以利用输入图像特征计算 PTC 损失（式（2）），结合历史信息指导弱监督语义分割。

为了进一步精细化记忆库损失的优化过程，本文增加了一个相似度系数 ρ 来自适应调整损失权值。具体而言，分别计算 $S_{\cos}$ 与最终 CAM $\boldsymbol{M}$ 以及辅助 CAM $\boldsymbol{M}^{\mathrm{mid}}$ 的相似度，通过 $\boldsymbol{S}_{\mathrm{CAM}} = \frac{\boldsymbol{S}_{\cos} \cdot \boldsymbol{M}}{\| \boldsymbol{S}_{\cos} \| \| \boldsymbol{M} \|}$ 以及 $\boldsymbol{S}_{\mathrm{midCAM}} = \frac{\boldsymbol{S}_{\cos} \cdot \boldsymbol{M}^{\mathrm{mid}}}{\| \boldsymbol{S}_{\cos} \| \| \boldsymbol{M}^{\mathrm{mid}} \|}$，分别生成两个概率分布图 $\boldsymbol{S}_{\mathrm{CAM}}$ 和 $\boldsymbol{S}_{\mathrm{midCAM}}$。当两种概率分布图相似度较高时，表明记忆库与 CAM 和辅助 CAM 都有很好的一致性，可以有效地指导图像分割。在这种情况下，应该增加损失权值；否则，减小权重。为了实现该策略，本文引入巴赫曼距离来度量两个分布图之间的相似度。首先，通过 L2 范数对概率图进行归一化，并设置 0～1 内的 10 个 bin 区间来计算两个概率图的概率直方图。然后，通过测量分布图之间巴赫曼距离 D_{B} 得到相似度系数 ρ，即：

$$\begin{aligned}\rho &= 1 - D_{\mathrm{B}}(h_{\mathrm{CAM}}, h_{\mathrm{midCAM}}) = \\ &1 - \left[-\ln\left(\sum_{i=1}^{10} \sqrt{h_{\mathrm{CAM}}(i) \cdot h_{\mathrm{midCAM}}(i)} \right) \right],\end{aligned} \tag{8}$$

其中：i 表示直方图中的某个区间，h_{CAM}，h_{midCAM} 分别代表 $\boldsymbol{S}_{\mathrm{CAM}}$ 以及 $\boldsymbol{S}_{\mathrm{midCAM}}$ 的直方图。基于上述设置，根据式（2），最终的时间上下文对比学习模块的损失 $\mathcal{L}_{\mathrm{tcc}}$ 设计如下：

$$\begin{aligned}\mathcal{L}_{\mathrm{tcc}} = &\frac{1}{2}\rho\left(1 - \frac{\sum_{i,j} \mathcal{M}^{+} \circ \mathrm{ASim}(F_i, F_j)}{N^{+}}\right) + \\ &\frac{1}{2}\rho\left(\frac{\sum_{i,j} \mathcal{M}^{-} \circ \mathrm{ASim}(F_i, F_j)}{N^{-}}\right).\end{aligned} \tag{9}$$

3.4 总损失

除了上文介绍的空间特征对比学习模块中的补丁级令牌对比损失 $\mathcal{L}_{\mathrm{ptc}}$ 及类级令牌对比损失 $\mathcal{L}_{\mathrm{ctc}}$、时间上下文对比学习损失 $\mathcal{L}_{\mathrm{mpc}}$ 外，总损失还包括主分类器 $\mathcal{L}_{\mathrm{cls}}$、辅助分类器的损失 $\mathcal{L}_{\mathrm{mcls}}$ 以及分割损失 $\mathcal{L}_{\mathrm{seg}}$。本文使用多标签分类损失计算分类损失，以主分类器为例，损失函数如下：

$$\begin{aligned}&\mathcal{L}_{\mathrm{cls}} = \\ &-\frac{1}{C}\sum_{i=1}^{C}\left[y_i \log\left(\sigma(\hat{y}_i)\right) + (1-y_i)\log\left(1-\sigma(\hat{y}_i)\right)\right],\end{aligned} \tag{10}$$

其中：C 为类别的数量，y_i 代表第 i 个类别的真实标签，$\hat{y}_i$ 为第 i 个类别的预测值，σ 表示 Sigmoid 激活函数。与之前的单阶段弱监督语义分割设置[32,36]一样，本文引入像素自适应细化模块（Property-Aware Relation，PAR）对伪标签进行细化，改进后的伪标签将用于监督分割解码器的训练。本文模型的总损失则是这些损失项的加权和，如下

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{ptc}} + \lambda_2 \mathcal{L}_{\mathrm{ctc}} + \lambda_3 \mathcal{L}_{\mathrm{tcc}} + \mathcal{L}_{\mathrm{cls}} + \mathcal{L}_{\mathrm{mcls}} + \mathcal{L}_{\mathrm{seg}}, \tag{11}$$

其中 λ_1，λ_2，λ_3 代表各损失的权重。

4 实　验

4.1 实验设置

4.1.1 数据集和评估指标

在 PASCAL VOC[44] 和 MS COCO[43] 数据集上评估了本文提出的 STCL 网络。PASCAL VOC 数据集共包含 21 个类（20 个目标类别和背

景），包括一个训练集（1 464 张图像），一个验证集（1 449 张图像）和一个测试集（1 456 张图像）。遵循传统的做法，VOC 数据集用 SBD 数据集[45]进行扩充，扩充后训练集包含 10 582 张图像。MS COCO 由 81 个类（80 个目标类别和背景）组成，其中 80 000 张图像用于训练，40 000 张图像用于验证。本文采用平均交并比（mean Intersection over Union，mIoU）评估 PASCAL VOC 和 MS COCO 数据集上生成的 CAM 和分割结果的精度。mIoU 计算了预测类别与实际类别之间的交集与并集的比值，用来评估预测值与实际值之间的相似程度，其计算公式如下：

$$mIoU=\frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}},\quad(12)$$

其中：k 为数据集中除背景的目标类别数目，i 表示真实值，j 表示预测值，p_{ij} 表示将真实值 i 预测成类别 j 的像素数量，mIoU 的取值是 0～1，数值越高证明精度越高。

4.1.2　实现细节

采用基于 ViT 框架的弱监督语义分割网络作为本文的基准网络。本文采用具有分类和分割头的 ViT 作为所提出方法的骨干网络，分类头是一个全连接网络，分割头由两个膨胀率为 5 的 3×3 卷积层和一个 1×1 预测层组成。该网络由 ImageNet 预训练模型进行初始化。使用学习率为 0.000 06、权衰减为 0.01 的 AdamW 优化器训练所提出的模型。按照之前工作的策略[18]，输入图像被增强并调整大小为 448×448，并在推理阶段使用多尺度测试和 CRF 后处理。

在 VOC 数据集上的实验，batch size 设为 4，迭代次数是 20 000。在空间特征对比学习模块中，全局和局部图像分别设置为 448×448 和 96×96，背景阈值 (α,β) 设为 $(0.25,0.7)$，式（3）中的温度因子设为 0.5，ε 设为 0.000 1。在时间上下文对比学习模块中，动量因子 γ 设置为 0.1，记忆库每 200 次迭代更新一次。在 4.2.3 节中，损失函数中的权值 $(\lambda_1,\lambda_2,\lambda_3)$ 设置为 $(0.2,0.5,0.5)$。

在 COCO 数据集上的实验，batch size 和迭代次数分别设置为 8 和 80 000。在空间特征对比学习模块中，背景阈值 (α,β) 设置为 $(0.25,0.65)$。其他参数依照 VOC 数据集进行设置。

4.2　实验结果

4.2.1　CAM 和伪标签

使用 VOC 数据集的训练集和验证集评估 CAM 结果，并将其与最近先进的弱监督语义分割方法进行比较。表 1 总结了由本文 CAM 生成的伪标签的评估结果，以及与其他方法的比较结果。由于 ViT 最初是在 ImageNet-21k 上进行预训练的，本文首先获得了 ImageNet-21k 权重的网络分割结果，记为 ViT-B†。为了与其他方法进行公平的比较，还基于 ImageNet-1k 预训练权值评估了 STCL 网络，记为 ViT-B。从表 1 可以看出，伪标签准确率在 VOC 训练集上达到 74.2%，在验证集上达到 72.7%。

表 1　VOC 数据集中 CAM 的评估结果

Tab. 1　Evaluation results of CAM on VOC dataset（%）

方　法	骨干网络	训练集	验证集
多阶段方法			
PPC[12]$_{\text{CVPR'2022}}$+PSA[29]	WR-38	73.3	—
ACR[33]$_{\text{CVPR'2023}}$+IRN[46]	WR-38	72.3	—
CTI[20]$_{\text{CVPR'2024}}$	ViT-B	73.7	—
SFC[47]$_{\text{AAAI'2024}}$	WR-38	73.7	—
单阶段方法			
RRM[35]$_{\text{AAAI'2020}}$	WR-38	—	65.4
1Stage[37]$_{\text{CVPR'2020}}$	WR-38	66.9	65.3
AA&LR[28]$_{\text{ACM MM'2021}}$	WR-38	68.2	65.8
SLRNet[36]$_{\text{IJCV'2022}}$	WR-38	67.1	66.2
AFA[40]$_{\text{CVPR'2022}}$	MiT-B1	68.7	66.5
ViT-PCM[39]$_{\text{ECCV'2022}}$	ViT-B†	67.7	66.0
ViT-PCM+CRF[29]$_{\text{ECCV'2022}}$	ViT-B†	71.4	69.3
ToCo[18]$_{\text{CVPR'2023}}$	ViT-B	72.2	70.5
ToCo†[18]$_{\text{CVPR'2023}}$	ViT-B†	73.6	72.3
MCC[48]$_{\text{WACV'2024}}$	ViT-B	73.0	71.3
RTC[49]$_{\text{TCSVT'2024}}$	WR-38	65.5	63.1
本文方法 STCL	ViT-B	73.1	71.5
本文方法 STCL†	**ViT-B†**	**74.2**	**72.7**

与其他基于ViT的方法相比，无论使用ImageNet-21k还是ImageNet-1k，本文方法都能生成比竞争对手更高质量的伪标签。此外，弱监督语义分割方法通常分为单阶段方法和多阶段方法，在单阶段方法中，伪标签是直接生成的，而多阶段方法通过更复杂的加工流程生成更准确的伪标签。从对比结果中得知，STCL可以超越最近的单阶段竞争对手，甚至超过多阶段方法。本文方法的CAM可视化结果如图3所示。

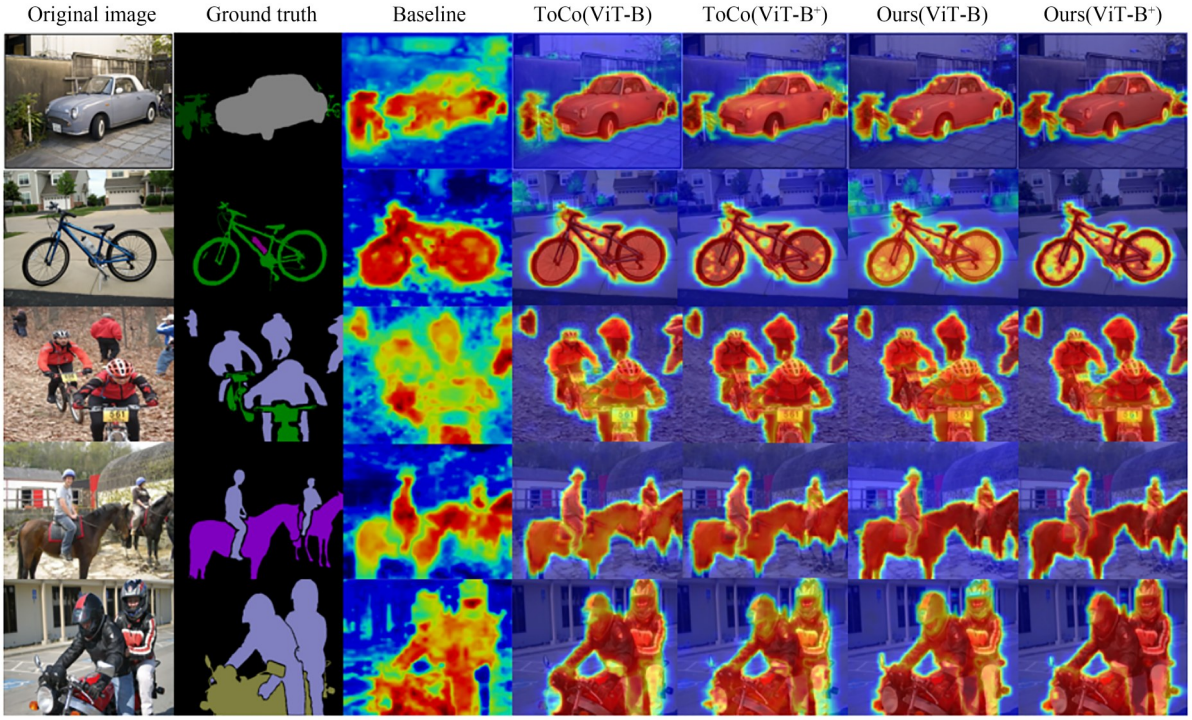

图3　VOC数据集上的CAM可视化结果

Fig. 3　Visualization results of CAM on VOC dataset

4.2.2　语义分割结果

为了证明所提出的模型的优越性，在表2中统计了本文方法以及最近先进弱监督语义分割方法在VOC和COCO数据集上的语义分割结果。从表中可以看出，STCL在VOC验证集上的mIoU值达到了72.7%，VOC测试集的准确率为73.1%，COCO验证集的准确率为43.6%，优于最近先进的单阶段弱监督语义分割方法。尽管一些多阶段弱监督语义分割方法使用了额外的监督信息，例如加入显著性图监督（$I+S$），并实现了较高的分割精度，但与之相比，STCL仍然表现出具有竞争力的分割性能。本文方法与其他方法对比的可视化结果如图4所示，可以看出，STCL在细节处理上表现出更优的效果。

由表2可知，不同的单阶段方法使用了不同的骨干网络。除了本文方法使用的ViT-B和ViT-B†外，表中的骨干网络还包括WR-38（Wide ResNet38）和MiT-B1（MixFormer-Base1）。为了消除骨干网络对分割性能的影响，利用这些方法在完全监督分割中的性能表现来评估它们，以进行公平的比较。比较结果如表3所示，当使用图像级标签作为监督时，所提出网络使用ViT-B达到了88.5%的全监督性能，显示了其相对于其他单阶段弱监督语义分割方法的竞争优势。特别地，本文模型优于其他基于ViT-B的模型，如ToCo[18]，RTC[49]。

4.2.3　消融实验

为了评估所提出方法的有效性，设计了一系列消融实验来评估STCL在VOC数据集上的组

表 2　VOC 和 COCO 数据集上的语义分割结果

Tab. 2　Semantic segmentation results on VOC and COCO datasets　　(%)

方　法	骨干网络	监督方式	VOC 验证集	VOC 测试集	COCO 验证集
多阶段方法					
RIB[23]$_{\text{NeurIPS'2021}}$	DL-V2	$\mathcal{I}+\mathcal{S}$	70.2	70.0	—
EPS[27]$_{\text{CVPR'2021}}$	DL-V2	$\mathcal{I}+\mathcal{S}$	71.0	71.8	—
L2G[24]$_{\text{CVPR'2022}}$	DL-V2	$\mathcal{I}+\mathcal{S}$	72.1	71.7	44.2
RCA[25]$_{\text{CVPR'2022}}$	DL-V2	$\mathcal{I}+\mathcal{S}$	72.2	72.8	36.8
PPC[12]$_{\text{CVPR'2022}}$	DL-V2	$\mathcal{I}+\mathcal{S}$	72.6	73.6	—
CLIP-ES[31]$_{\text{CVPR'2023}}$	DL-V2	$\mathcal{I}+\mathcal{T}$	71.1	71.4	45.4
VWL[32]$_{\text{IJCV'2022}}$	DL-V2	$\mathcal{I}$	69.2	69.2	36.2
W-OoD[30]$_{\text{CVPR'2022}}$	WR-38	$\mathcal{I}$	70.7	70.1	—
MCTformer[16]$_{\text{CVPR'2022}}$	WR-38	$\mathcal{I}$	71.9	71.6	42.0
ESOL[22]$_{\text{NeurIPS'2022}}$	DL-V2	$\mathcal{I}$	69.9	69.3	42.6
OCR[34]$_{\text{CVPR'2023}}$	WR-38	$\mathcal{I}$	72.7	72.0	42.5
单阶段方法					
RRM[35]$_{\text{AAAI'2020}}$	WR-38	$\mathcal{I}$	62.6	62.9	—
1Stage[37]$_{\text{CVPR'2020}}$	WR-38	$\mathcal{I}$	62.7	64.3	—
AFA[40]$_{\text{CVPR'2022}}$	MiT-B1	$\mathcal{I}$	66.0	66.3	38.9
SLRNet[36]$_{\text{IJCV'2022}}$	WR-38	$\mathcal{I}$	67.2	67.6	35.0
TSCD[41]$_{\text{AAAI'2023}}$	MiT-B1	$\mathcal{I}$	67.3	67.5	40.1
ToCo[18]$_{\text{CVPR'2023}}$	ViT-B	$\mathcal{I}$	69.8	70.5	41.3
ToCo†[18]$_{\text{CVPR'2023}}$	ViT-B†	$\mathcal{I}$	71.1	72.2	42.3
MCC[48]$_{\text{WACV'2024}}$	ViT-B	$\mathcal{I}$	70.3	71.2	42.3
RTC[49]$_{\text{TCSVT'2024}}$	WR-38	$\mathcal{I}$	67.2	68.8	35.1
RTC[49]$_{\text{TCSVT'2024}}$	ViT-B†	$\mathcal{I}$	72.4	73.0	42.5
本文方法 STCL	ViT-B	$\mathcal{I}$	71.3	71.4	42.5
本文方法 STCL†	**ViT-B†**	$\mathcal{I}$	**72.7**	**73.1**	**43.6**

Note:"$\mathcal{I}$" denotes image-level label supervision, "$\mathcal{S}$" denotes saliency map supervision, "$\mathcal{T}$" denotes text-driven supervision

表 3　弱监督分割模型与全监督模型的分割性能比较(全监督采用像素级伪标签对分割头进行监督)

Tab. 3　Performance comparison between weakly supervised and fully supervised segmentation models and fully-supervised models use pixel-level pseudo labels to supervise segmentation head　　(%)

方　法	骨干网络	VOC 验证集(F)	VOC 验证集(I)	比例(ratio)
1Stage[37]$_{\text{CVPR'2020}}$	WR-38	80.8	62.7	77.6
SLRNet[36]$_{\text{IJCV'2022}}$	WR-38	80.8	67.2	83.2
AFA[40]$_{\text{CVPR'2022}}$	MiT-B1	78.7	66.0	83.9
ToCo[18]$_{\text{CVPR'2023}}$	ViT-B	80.5	69.8	86.7
ToCo†[18]$_{\text{CVPR'2023}}$	ViT-B†	82.3	71.1	86.4
本文方法 STCL	ViT-B	80.6	71.3	88.5
本文方法 STCL†	**ViT-B†**	**82.5**	**72.7**	**88.1**

Note:"F" denotes pixel-level supervision, "I" denotes image-level supervision, ratio=val(I)/val(F)

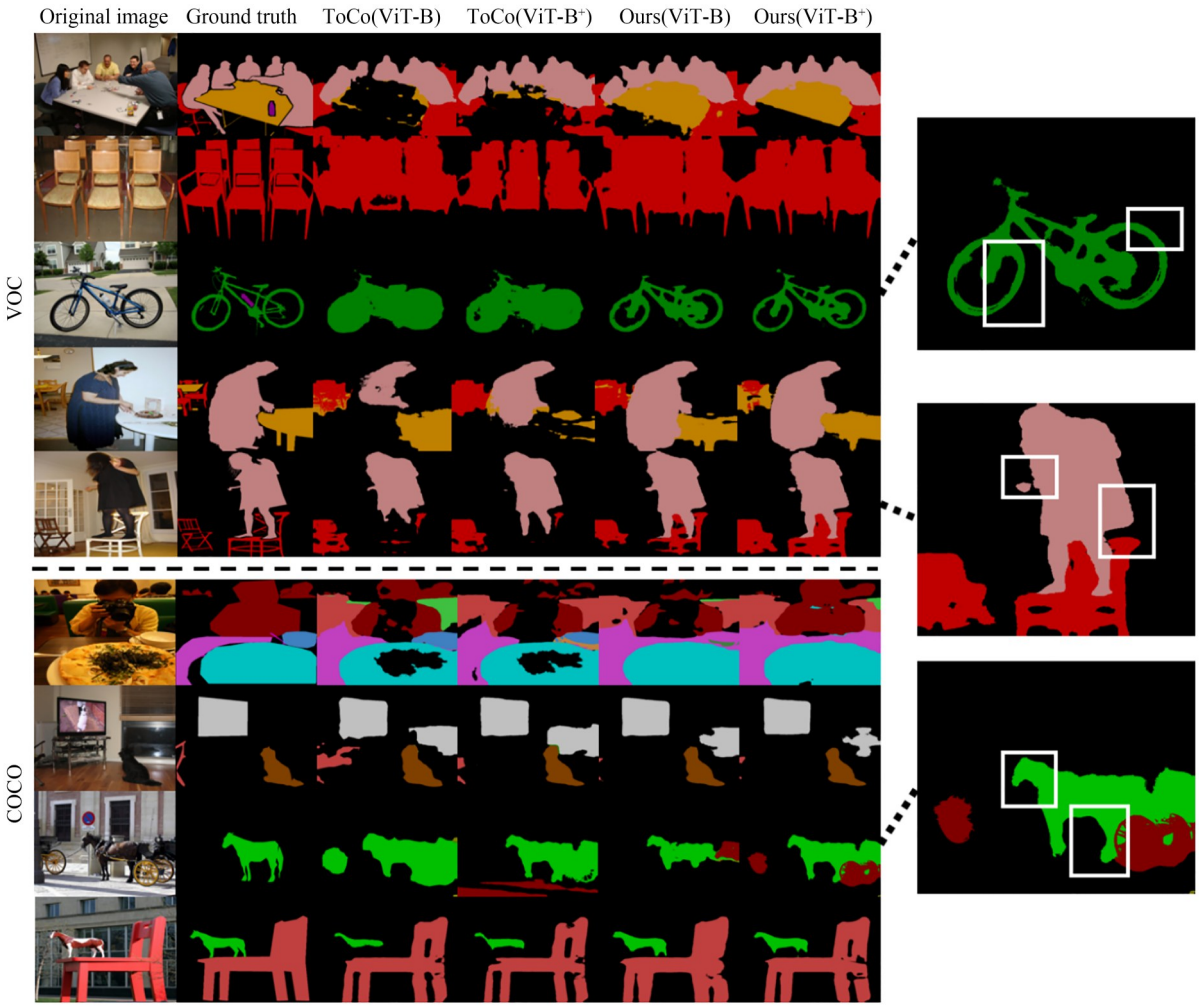

图 4　VOC 和 COCO 数据集上的语义分割可视化结果

Fig. 4　Semantic segmentation visualization results on VOC and COCO datasets

件性能和参数设置。特别地，消融实验均未使用 CRF 后处理，且 STCL 网络是在 ImageNet-1k 上预训练的。

4.2.3.1　组件消融

本文模型主要由空间特征对比学习和时间上下文对比学习两个模块组成，表 4 给出了 STCL 在 VOC 数据集上不同组件组合的 CAM 和分割精度结果。从结果可以看出，与基准网络相比，空间特征对比学习模块充分利用了令牌的语义相关性，CAM 和分割结果的 mIoU 分别提高了 6.0% 和 3.7%。基于强大的历史信息提取能力，时间上下文对比学习模块将 CAM 和分割的 mIoU 分别提高了 4.3% 和 2.9%。当这两个模块一起应用时，分割效果最佳，CAM 和分割的 mIoU 分别达到 71.5% 和 69.0%，证明了本文网络的有效性。STCL 组件消融的可视化结果如图 5 所示。可以看出，在两个模块中，空间特征对比学习对模型精度的贡献较大，其次是时间上下文对比学习模块。

表 4　不同模块组合在 VOC 验证集上的性能比较

Tab. 4　Performance comparison of different module combinations on VOC validation set　(%)

基准	空间特征对比学习	时间上下文对比学习	CAM 精度	分割精度
√			61.2	60.3
√	√		67.2	64.0
√		√	65.5	63.2
√	**√**	**√**	**71.5**	**69.0**

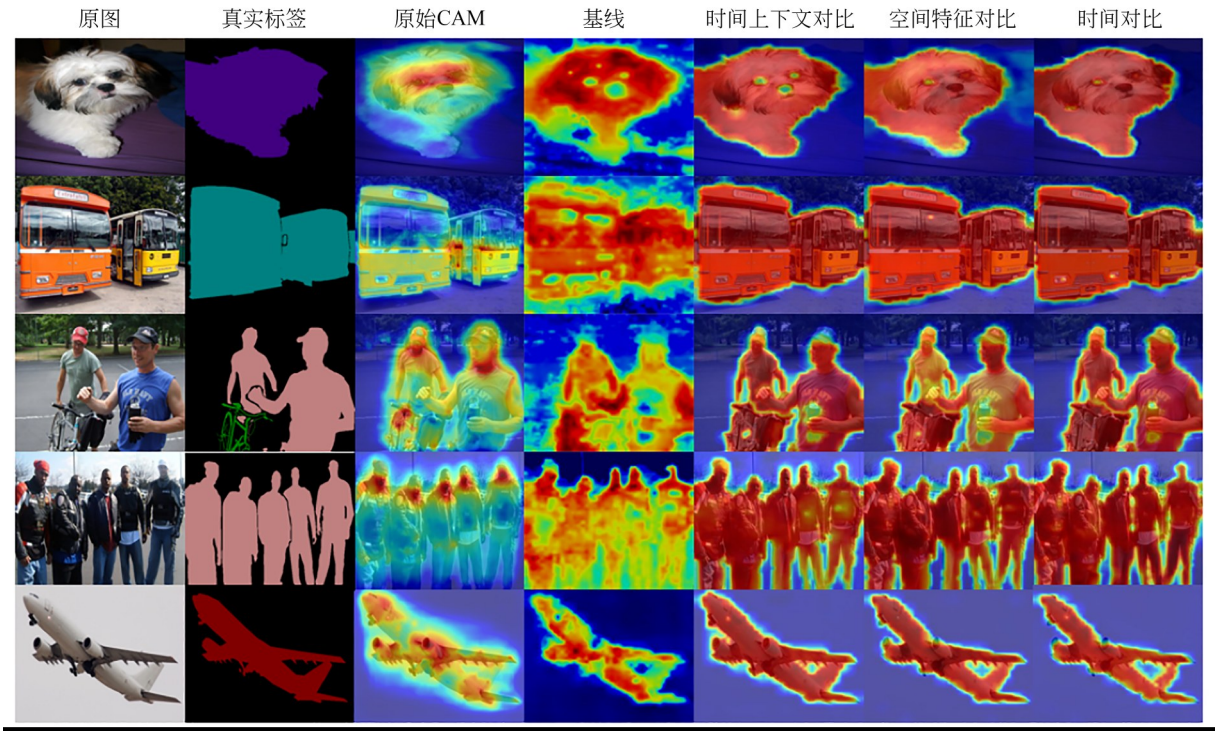

图 5　不同模块配置下的CAM可视化结果

Fig. 5　CAM visualization results under different module configurations

4.2.3.2　时间上下文对比学习中的更新频率

在构建记忆库时,随着新图像的输入,记忆库会不断更新,但是由于图像中存在噪声,如果每张图像都用于更新记忆库,不仅不会增加记忆库的准确性,反而带来反效果。为了探究记忆库更新频率对STCL的影响,本文分析了不同更新频率下VOC数据集上的分割结果,如表5所示。可以看出,当时间上下文对比学习模块中的更新频率设置为200次迭代时,CAM和分割精度均可达到最佳性能。

表5　时间上下文对比学习模块中不同更新频率下的mIoU结果

Tab. 5　Results of mIoU with different update frequencies in temporal context contrastive learning module

(%)

更新频次	CAM精度	分割精度
50	70.6	68.8
100	71.0	68.7
200	**71.5**	**69.0**
400	71.1	68.8
800	70.4	68.3

4.2.3.3　时间上下文对比学习中的动量因子

在记忆库更新过程中,设置了动量因子γ控制更新的平滑程度,进而保证更新的稳定。赋予了动量因子不同的阈值,在VOC数据集上测试动量因子在不同阈值情况下的分割精度的变化。如表6所示,当动量因子设置为0.1时,记忆库达到了最优性能,STCL模型的分割精度可以达到69.0%。

表6　时间上下文对比学习模块中不同动量因子下的mIoU结果

Tab. 6　Results of mIoU with different momentum factors in temporal context contrastive learning module

(%)

动量因子	CAM精度	分割精度
0.01	70.4	68.5
0.05	71.2	68.9
0.10	**71.5**	**69.0**
0.20	71.1	68.8
0.30	70.8	68.6
0.50	70.3	68.3

4.2.3.4　空间特征对比学习中的辅助分类器

在空间特征对比学习模块中,辅助分类器不仅提供了辅助分类损失,而且通过辅助CAM与

最终CAM进行了比较，为STCL贡献了补丁级令牌对比损失。当使用ViT-B作为骨干网络时，网络中共有12个Transformer块，通常情况下，浅层块无法捕获高级语义，而后期块则会遇到过度相似的问题，因此需要设计实验，选择最优块来添加辅助分类器。表7展示了使用不同的Transformer块对生成CAM的精度和分割精度的影响。从表中可以看出，当选择第10个Transformer块时，STCL在CAM和分割精度方面的性能都是最优的。因此，本文选择第10块Transformer块添加辅助分类器，并利用辅助CAM在STCL中监督最终CAM的生成。

表7　不同辅助分类器在空间特征对比学习中的结果

Tab. 7　Performance results of different auxiliary classifiers in spatial feature contrastive learning　(%)

Transformer块	CAM精度	分割精度
第8块	64.1	61.2
第9块	68.7	65.7
第10块	**71.5**	**69.0**
第11块	43.8	40.8

4.2.3.5　背景阈值设置

在空间特征对比学习模块中，设置了背景阈值(α,β)将CAM分为背景、前景和不确定区域。利用mIoU测试了不同的背景阈值对于分割精度的影响，结果如表8所示。从表中可以看出，对于VOC数据集，当背景阈值(α,β)设置为(0.25, 0.7)时，STCL的分割性能最优，mIoU可以达到69.0%。

表8　不同背景阈值下的语义分割结果比较

Tab. 8　Comparison of semantic segmentation results under different background thresholds　(%)

β	α			
	0.15	0.2	**0.25**	0.3
0.60	—	55.9	63.8	67.6
0.65	51.9	62.3	68.1	67.2
0.70	57.0	66.4	**69.0**	68.0
0.75	63.9	66.5	67.2	65.2

4.2.4　应用实验

为了评估本文方法在不同领域的有效性，测试了STCL在医学图像和工业图像场景中的分割性能。与自然场景图像相比，医学图像具有细微的类间差异、对上下文依赖强及边界模糊且结构密集等特征。工业图像的特征则是纹理模式复杂、缺陷形态多样且目标区域往往呈现出低对比度和不规则形状。这些固有的复杂性使得弱监督语义分割在这些领域的应用具有一定的挑战性。

具体来说，这里评估了STCL在PanNuke数据集[50]以及NEU-Seg数据集[51]的分割性能。PanNuke是一个用于细胞核实例分割的半自动生成的数据集，包括5个类别和205 343个带标签的细胞核，每个细胞核都附有一个实例分割掩码。将每种类型按8∶2的比例划分为训练集和测试集后，最终得到6 312张图像组成的训练集和1 589张图像组成的测试集。NEU-Seg数据集是一种钢材表面缺陷数据集，包含三类缺陷：夹渣、斑块、划痕。数据集共包含3 630张训练图像和840张测试图像，其中测试图像包含600张单一缺陷类别图像和240张多缺陷类别图像。

使用mIoU评估STCL在医学图像和工业图像场景中的性能，结果如表9所示。从表中可以看出，STCL的分割结果优于其他先进的弱监督语义分割方法。在这两种场景下的可视化分割结果如图6所示。与其他方法相比，不管是医学图像还是工业缺陷图像，STCL都生成了更精确的物体边界，更好地接近真实标签。这证明了所提出的方法在实际应用中的可行性和有效性。

表9　其他应用场景图像中的语义分割比较结果

Tab. 9　Semantic segmentation comparison results in images from other application scenarios　(%)

方　法	PanNuke数据集	NEU-Seg数据集
1Stage[37]	32.5	50.2
SLRNet[36]	39.4	56.5
AFA[40]	42.9	60.5
ToCo[18]	47.0	65.4
STCL	**49.2**	**68.8**

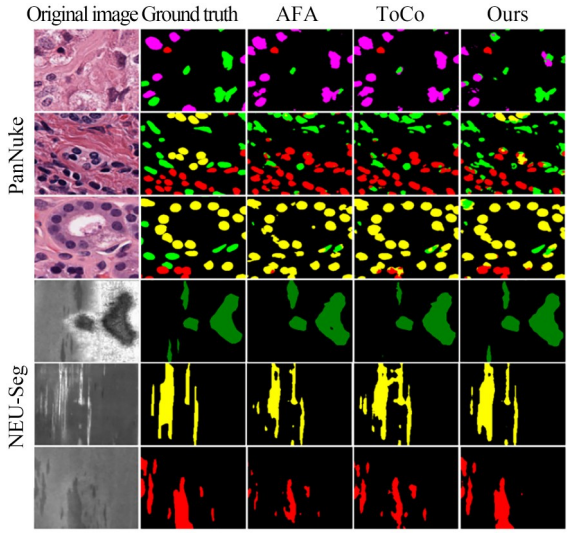

图 6　不同应用场景中语义分割可视化结果比较

Fig. 6　Comparison of semantic segmentation visualization results across different application scenarios

5　结　论

本文提出了一种时空对比学习驱动的弱监督图像语义分割网络，该网络通过从时间和空间维度挖掘图像中隐含的监督信息，有效提升了弱监督语义分割的精度。具体来说，利用 ViT 的令牌机制，引入了空间特征对比学习模块，通过补丁级和类级令牌的对比策略，深入探索图像空间中的语义特征关系，从而优化 CAM 的生成。此外，提出了一种时间上下文对比学习模块，通过构建记忆库引入历史先验知识为 CAM 的进一步优化提供指导，并结合记忆库更新策略和自适应记忆对比损失函数，进一步增强了模型对细节区域的辨识能力。实验结果表明，STCL 在 PASCAL VOC 和 MS COCO 数据集上均取得了优异的性能，平均交并比可以分别达到 72.7% 以及 43.6%，达到了 88.5% 的全监督性能，超越了先进的多阶段和单阶段弱监督语义分割方法。未来的研究将进一步探索模型在不同领域、不同模态下的泛化能力，从而提升模型在多源弱监督任务中的普适性与迁移性能。

作者贡献声明：

梁臻：方法提出，论文构思与撰写；

胡燕祝：数据管理与分析；

杨洋：可视化与论文审核。

参考文献：

[1] 李智杰，惠爱婷，李昌华，等. 面向遥感图像道路提取的多尺度上下文感知网络[J]. 光学 精密工程，2025，33(4)：610-623.
LI ZH J，HUI A T，LI CH H，*et al*. Multi-scale context-aware network for road extraction in remote sensing images[J]. *Optics and Precision Engineering*，2025，33(4)：610-623.（in Chinese）

[2] 徐晗晗，张印辉，何自芬，等. 伪标签置信度调控结直肠癌病理图像半监督语义分割[J]. 光学 精密工程，2025，33(4)：591-609.
XU H H，ZHANG Y H，HE Z F，*et al*. Pseudo-label confidence regulates semi-supervised semantic segmentation of pathological images of colorectal cancer [J]. *Optics and Precision Engineering*，2025，33(4)：591-609.（in Chinese）

[3] 李文生，张菁，卓力，等. 基于Transformer的视觉分割技术进展[J]. 计算机学报，2024，47(12)：2760-2782.
LI W SH，ZHANG J，ZHUO L，*et al*. Overview of transformer-based visual segmentation techniques [J]. *Chinese Journal of Computers*，2024，47(12)：2760-2782.（in Chinese）

[4] 田萱，王亮，丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报，2019，30(2)：440-468.
TIAN X，WANG L，DING Q. Review of image semantic segmentation based on deep learning[J]. *Journal of Software*，2019，30(2)：440-468.（in Chinese）

[5] 李军侠，苏京峰，崔滢，等. 基于语义调制的弱监督语义分割[J]. 软件学报，2025，36(9)：4373-4387.
LI J X，SU J F，CUI Y，*et al*. Semantic-modulation-based weakly supervised semantic segmentation [J]. *Journal of Software*，2025，36(9)：4373-4387.（in Chinese）

[6] LEE J，YI J H，SHIN C，*et al*. BBAM：bounding box attribution map for weakly supervised semantic and instance segmentation[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 20-25，2021，*Nashville*，*TN*，

USA. IEEE, 2021: 2643-2651.

[7] OH Y, KIM B, HAM B. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation[C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021. *Nashville*, *TN*, *USA*. IEEE, 2021: 6909-6918.

[8] BEARMAN A, RUSSAKOVSKY O, FERRARI V, *et al*. *What's the Point*: *Semantic Segmentation with Point Supervision*[M]. Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 549-565.

[9] LIN D, DAI J F, JIA J Y, *et al*. ScribbleSup: scribble-supervised convolutional networks for semantic segmentation[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 3159-3167.

[10] VERNAZA P, CHANDRAKER M. Learning random-walk label propagation for weakly-supervised semantic segmentation[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *July* 2017, *Honolulu*, *HI*, *USA*. IEEE, 2017: 2953-2961.

[11] DU Y, FU Z H, LIU Q J, *et al*. Weakly supervised semantic segmentation by pixel-to-prototype contrast [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 4310-4319.

[12] ZHOU B L, KHOSLA A, LAPEDRIZA A, *et al*. Learning deep features for discriminative localization[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 2921-2929.

[13] WU Y C, LI X Q, DAI S M, *et al*. Hierarchical semantic contrast for weakly supervised semantic segmentation[C]. *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*. *August* 19-25, 2023, *SAR*, *China*. 2023: 1542-1550.

[14] SUN W, ZHANG J, LIU Z, *et al*. Getam: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation [J/OL]. *arXiv*, (2021-12-6). https://arxiv.org/abs/2112.02841.

[15] WU W Y, DAI T H, HUANG X W, *et al*. Top-K pooling with patch contrastive learning for weakly-supervised semantic segmentation [C]. 2024 *IEEE International Conference on Systems*, *Man*, *and Cybernetics* (*SMC*). *October* 6-10, 2024, *Kuching*, *Malaysia*. IEEE, 2024: 5270-5275.

[16] XU L, OUYANG W L, BENNAMOUN M, *et al*. Multi-class token transformer for weakly supervised semantic segmentation [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 4300-4309.

[17] XU L, BENNAMOUN M, BOUSSAID F, *et al*. MCTformer+: multi-class token transformer for weakly supervised semantic segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(12): 8380-8395.

[18] RU L X, ZHENG H L, ZHAN Y B, *et al*. Token contrast for weakly-supervised semantic segmentation [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 3093-3102.

[19] SUN W X, ZHANG Y H, QIN Z, *et al*. All-pairs consistency learning for weakly supervised semantic segmentation[C]. 2023 *IEEE/CVF International Conference on Computer Vision Workshops* (*ICCVW*). *October* 2-6, 2023, *Paris*, *France*. IEEE, 2023: 826-837.

[20] YOON S H, KWON H, KIM H, *et al*. Class tokens infusion for weakly supervised semantic segmentation [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 3595-3605.

[21] CHEN Z Z, WANG T, WU X W, *et al*. Class reactivation maps for weakly-supervised semantic segmentation [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 959-968.

[22] LI J L, JIE Z Q, WANG X, *et al*. Expansion and shrinkage of localization for weakly-supervised semantic segmentation[C]. *Advances in Neural Information Processing Systems*, *New Orleans*, *USA*. 2022, 35: 16037-16051.

[23] LEE J, CHOI J, MOK J, *et al*. Reducing informa-

tion bottleneck for weakly supervised semantic segmentation [C]. *Advances in Neural Information Processing Systems*, 2021, 34: 27408-27421.

[24] JIANG P T, YANG Y Q, HOU Q B, *et al*. L2G: a simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *OrleansNew*, *LA*, *USA*. IEEE, 2022: 16865-16875.

[25] ZHOU T F, ZHANG M J, ZHAO F, *et al*. Regional semantic contrast and aggregation for weakly supervised semantic segmentation [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 4289-4299.

[26] YANG Z W, FU K X, DUAN M H, *et al*. Separate and conquer: decoupling co-occurrence *via* decomposition and representation for weakly supervised semantic segmentation [C]. 2024 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 16-22, 2024, *Seattle*, *WA*, *USA*. IEEE, 2024: 3606-3615.

[27] LEE S, LEE M, LEE J, *et al*. Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation [C]. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 20-25, 2021. *Nashville*, *TN*, *USA*. IEEE, 2021: 5491-5501.

[28] ZHANG X R, PENG Z L, ZHU P, *et al*. Adaptive affinity loss and erroneous pseudo-label refinement for weakly supervised semantic segmentation [C]. *Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event China*. ACM, 2021: 5463-5472.

[29] AHN J, KWAK S. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition. June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 4981-4990.

[30] LEE J, OH S J, YUN S, *et al*. Weakly supervised semantic segmentation using out-of-distribution data [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 16876-16885.

[31] LIN Y Q, CHEN M H, WANG W X, *et al*. CLIP is also an efficient segmenter: a text-driven approach for weakly supervised semantic segmentation[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 15305-15314.

[32] RU L X, DU B, ZHAN Y B, *et al*. Weakly-supervised semantic segmentation with visual words learning and hybrid pooling[J]. *International Journal of Computer Vision*, 2022, 130(4): 1127-1144.

[33] KWEON H, YOON S H, YOON K J. Weakly supervised semantic segmentation *via* adversarial learning of classifier and reconstructor[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 11329-11339.

[34] CHENG Z S, QIAO P C, LI K H, *et al*. Out-of-candidate rectification for weakly supervised semantic segmentation [C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 23673-23684.

[35] ZHANG B F, XIAO J M, WEI Y C, *et al*. Reliability does matter: an end-to-end weakly supervised semantic segmentation approach [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(7): 12765-12772.

[36] PAN J W, ZHU P F, ZHANG K H, *et al*. Learning self-supervised low-rank network for single-stage weakly and semi-supervised semantic segmentation[J]. *International Journal of Computer Vision*, 2022, 130(5): 1181-1195.

[37] ARASLANOV N, ROTH S. Single-stage semantic segmentation from image labels [C]. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 13-19, 2020. *Seattle*, *WA*, *USA*. IEEE, 2020: 4252-4261.

[38] GAO W, WAN F, PAN X J, *et al*. TS-CAM: token semantic coupled attention map for weakly supervised object localization [C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 10-17, 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 2866-2875.

[39] ROSSETTI S, ZAPPIA D, SANZARI M, *et al*. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation [C]. *Computer Vision-ECCV* 2022. *Cham*: *Springer Nature Switzerland*, 2022: 446-463.

[40] RU L X, ZHAN Y B, YU B S, *et al*. Learning affinity from attention: end-to-end weakly-supervised semantic segmentation with transformers [C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *OrleansNew*, *LA*, *USA*. IEEE, 2022: 16825-16834.

[41] XU R T, WANG C W, SUN J X, *et al*. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(3): 3045-3053.

[42] CARON M, TOUVRON H, MISRA I, *et al*. Emerging properties in self-supervised vision transformers [C]. 2021 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). 2021, *Montreal*, *QC*, *Canada*. IEEE, 2021: 9630-9640.

[43] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: common objects in context [C]. *Computer Vision-ECCV* 2014. *Cham*: *Springer International Publishing*, 2014: 740-755.

[44] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, *et al*. The pascal visual object classes (VOC) challenge [J]. *International Journal of Computer Vision*, 2010, 88(2): 303-338.

[45] HARIHARAN B, ARBELÁEZ P, BOURDEV L, *et al*. Semantic contours from inverse detectors [C]. 2011 *International Conference on Computer Vision*. *November* 6-13, 2011, *Barcelona*, *Spain*. IEEE, 2011: 991-998.

[46] AHN J, CHO S, KWAK S. Weakly supervised learning of instance segmentation with inter-pixel relations [C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2019: 2204-2213.

[47] ZHAO X Q, TANG F L, WANG X Y, *et al*. SFC: shared feature calibration in weakly supervised semantic segmentation [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(7): 7525-7533.

[48] WU F W, HE J X, YIN Y F, *et al*. Masked collaborative contrast for weakly supervised semantic segmentation [C]. 2024 *IEEE/CVF Winter Conference on Applications of Computer Vision* (*WACV*). *January* 3-8, 2024, *Waikoloa*, *HI*, *USA*. IEEE, 2024: 851-860.

[49] WANG C Y, ZHANG D, YAN R. Boosting weakly-supervised image segmentation *via* representation, transform, and compensator [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024, 34(11): 11013-11025.

[50] GAMPER J, ALEMI KOOHBANANI N, BENET K, *et al*. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification [C]. *Digital Pathology*. *Cham*: *Springer*, 2019: 11-19.

[51] DONG H W, SONG K C, HE Y, *et al*. PGA-net: pyramid feature fusion and global context attention network for automated surface defect detection [J]. *IEEE Transactions on Industrial Informatics*, 2020, 16(12): 7448-7458.

作者简介：

梁　臻(1997—),女,山东泰安人,博士研究生,2022年于济南大学获得计算机硕士学位,主要研究方向为图像处理、计算机视觉和深度学习。E-mail: liangzhen@bupt.edu.cn

通讯作者：

胡燕祝(1970—),男,北京海淀人,博士,教授,博士生导师,2005年于北京交通大学获得博士学位,主要从事机器视觉、模式识别方面的研究。E-mail: bupt_automation_safety_yzhu@bupt.edu.cn