Multi-class Token Transformer forWeakly Supervised Semantic Segmentation

摘要

MCTformer 这篇论文是多阶段的弱监督语义分割任务，也是早期引入 Transformer 到弱监督语义分割任务的工作，该论文提出基于 Transformer 模型改进的 MCTformer-V1 及 MCTformer-V2 模型。MCTformer-V1 使用多个类的 token 来学习类的信息，利用 Transformer 注意力生成特定类别的目标定位图，并采用类感知训练策略进行训练。MCTformer-V2 在 MCTformer-V1 的基础上整合了 CAM 模块，并通过加强类标记之间的一致性来提高模型学习效率。本文在复现这篇论文的工作基础上，提出两个改进方案，其一是引入了 U-Net 网络来提取图像的全局特征，减少 Transformer 注意力不可避免地丢失非判别性特征的问题；其二是对亲合度矩阵进行转置变换以增强 CAM 细化能力。实验结果表明，本文复现工作基本达到这篇论文指标，且两个改进方案均得到一定的提升，表明所提方案的有效性。

关键词： MCTformer；弱监督语义分割任务；U-Net；亲合度矩阵；CAM

1 引言

弱监督语义分割是计算机视觉领域的重要研究方向之一，旨在通过利用弱监督信息来减少对像素级真实标签的依赖，像素级标签提供了简单的弱标签，仅指示某些类别的存在或缺失，但没有提供任何真实的定位信息。弱监督语义分割的一般流程包括三个步骤：1) 使用像素级标签训练多标签分类模型；2) 提取每个类别的类激活图（CAM）以生成 0-1 掩码，然后进一步细化以生成伪掩码；3) 将所有类别的伪掩码作为伪标签，以完全监督的方式训练语义分割模型。其中，最关键的步骤是如何生成高质量的 CAM。

弱监督语义分割任务分为单阶段与多阶段的任务，MCTformer 是目前多阶段任务中相对较新的使用 Transformer 模型的研究成果，也是目前本人研究关注的方向。这篇论文是基于 ViT 模型的改进，也是早期将 ViT 模型引入到弱监督语义分割任务中较好的研究成果，具有较高的研究价值，因此选择复现该论文，并对其进行改进。

2 相关工作

大多数现有的弱监督语义分割方法都依赖于 CAM [15] 从 CNN 中提取目标定位图。原始的 CAM 存在目标不完整及边界粗糙的问题，因此无法为语义分割网络的学习提供足够的

监督。为解决这一问题，有人提出了特定的分割损失来弥补分割监督的不足，包括 SEC 损失 [5]、CRF 损失 [8,12] 和对比损失 [4]。此外，还有一些研究侧重于改进从 CAM 中获得的伪标签。一些研究者利用子类别 [2] 和跨图像语义 [3,7,16] 来定位更准确的目标区域。为了解决标准分类目标损失函数的局限性，有人提出了正则化损失 [10,14] 来引导网络发现更多的物体区域。此外还有一些工作侧重于学习成对语义亲和度以完善 CAM 。Ahn 等人 [1] 提出了 AffinityNet 从原始 CAM 中学习相邻像素之间的亲和度关系，可以预测一个亲和度矩阵，通过随机行走的方式生成 CAM 。在 [10,13] 中，亲和度直接从分类网络的特征图中学习，以完善 CAM 图。此外，Xu 等人 [11] 提出了一种跨任务亲和度，这种亲和度是从弱分类网络中的显著性表征和分割表征中学习的。在弱监督多任务框架下，从显著性和分割表征中学习跨任务亲和力。监督的多任务框架中从突出和分割表征中学习。与之前基于 CNN 的弱监督语义分割方法相比，这篇论文提出了一种基于 Transformer 的模型，以提取特定类别的目标定位图，利用自注意力机制中的 Transformer 注意力图来进一步生成目标定位图。

3 本文方法

3.1 本文方法概述

这篇论文首先提出了一种基于 Transformer 的新型框架（MCTformer-V1），以利用来自 Transformer 注意力的特定类别目标定位图。MCTformer-V1 使用多个类 token，这些类 token 与嵌入位置信息的 patch token 拼接形成 Transformer 编码器的输入 token。在训练时采用了类感知训练策略，以确保不同的类标记能学习到不同的类特定表征，由类标记直接产生的类得分与真实类标签之间计算分类损失。MCTformer-V1 网络结构如图 1所示：

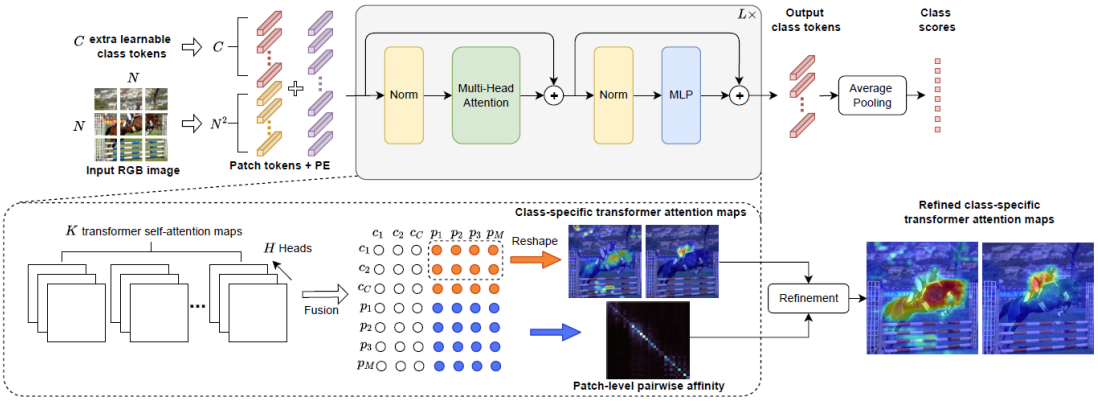

图 1. MCTformer-V1 网络结构图

此外，作者将 CAM 模块整合到所提出的 MCTformer-V1 框架中，构建一个扩展的 MCTformer-V2 模型，其中的 CAM 方法可以灵活、稳健地适应多标签图像。通过对来自类 token 和 patch token 的类预测上应用分类损失，这两类标记之间的一致性可以得到加强，从而提高模型学习的效率。MCTformer-V1 网络结构如图 2所示。

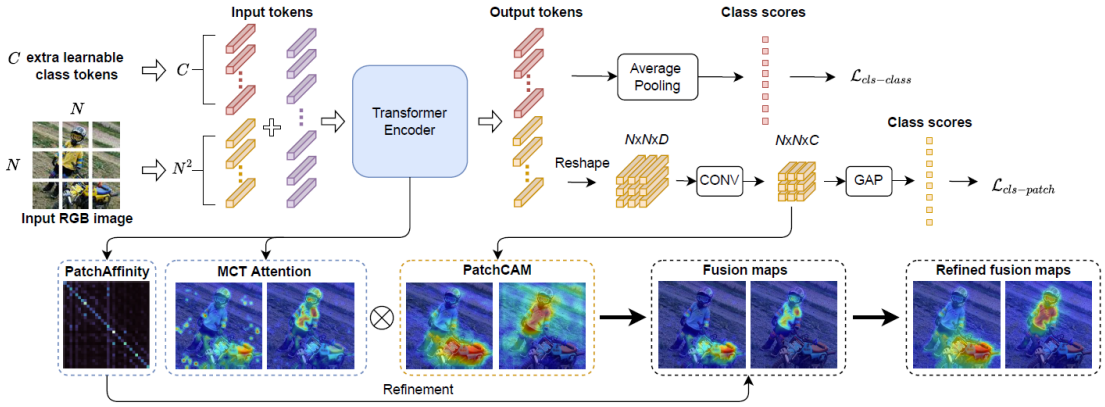

图 2. MCTformer-V2 网络结构图

3.2 U-Net 全局特征提取模块

这篇论文尽管使用了较为先进的 Transformer 模型来提取图像特征，但这种特征大多为判别性特征，仅包括目标对象的一些重要部位信息，而一些边缘的非判别性信息则不可避免地被缺失。因此为了能够减少这种信息丢失，尝试引入我所熟悉的 U-Net 网络 [6] 来提取图像的全局特征，加入从 Transformer 模型中提取到的类与类之间的特征到 U-Net 模型的每一层，使得模型能够学习分辨不同类之间的特征信息。U-Net 特征提取模块如图 3所示：

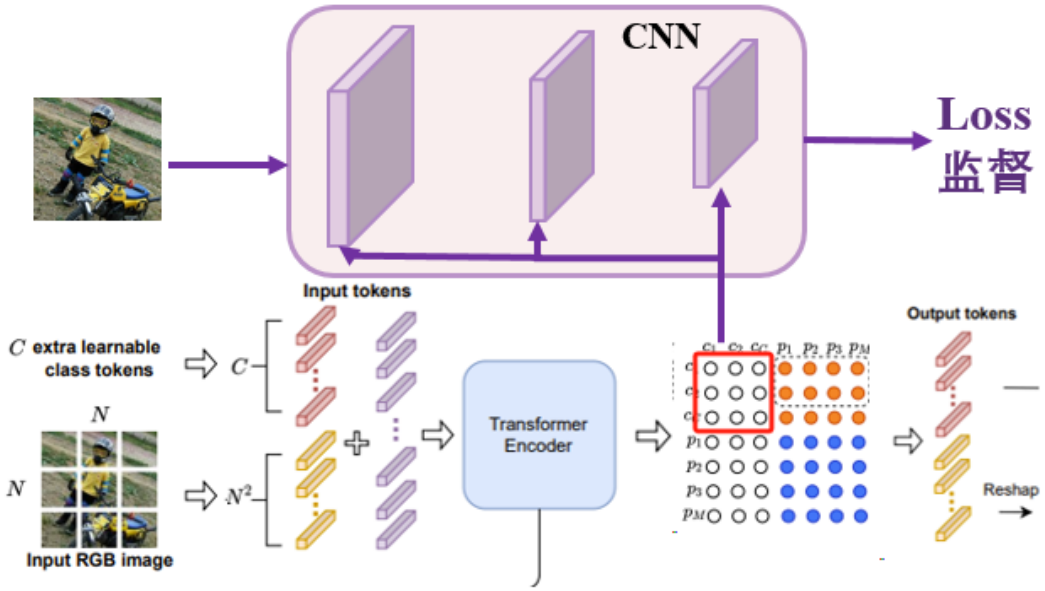

图 3. U-Net 特征提取模块

3.3 转置亲合度矩阵

本质上讲，自注意力机制是一种有向图模型 [9]，而多头注意力则是一种无向图模型，亲和度矩阵应该是对称的，因为共享相同语义的节点应该是相等的。因此我提出对亲合度矩阵进行转置变换，将亲合度 S 与其转置 S^T 相加即可。因此预测的亲合度矩阵 A 应为：

$$A = S + S^T \tag{1}$$

4 复现细节

4.1 与已有开源代码对比与创新点

本论文提供了源代码，但是存在部分细节及参数没有公开的问题，需要对代码进行调通，同时配置参数。因此，我先对论文进行复现工作，下载模型提供的权重及数据集，并对数据集进行论文所述的扩充处理，补充代码细节并配置参数，多次实验选择最优值。

此外，在成功复现该论文的基础上，我也对模型进行改进，进行了两个改进方案，方案一是加入3.2章节所述的 U-Net 全局特征提取模块；方案二是对论文中原有的模块进行改进，如3.3章节所述。

4.2 实验环境搭建

实验设置：训练集 VOC2012 扩充的 10582 张图像；验证集：1449 张图像；模型：MCTformer-V2 模型、MCTformer-V2+U-Net 模型、MCTformer-V2 改进模型；使用 pytorch 框架。

实验参数配置如图 4， 5和 6所示：

```
CUDA_VISIBLE_DEVICES=3,4,5 python main.py  --model deit_small_MCTformerV2_patch16_224  \
    --batch-size 64    \
    --data-set VOC12  \
    --img-list voc12   \
    --data-path Datasets/VOC2012  \
    --layer-index 12  \
    --output_dir MCTformer_results/voc/MCTformer_v2  \
    --finetune https://dl.fbaipublicfiles.com/deit/deit_small_patch16_224-cd65a155.pth
######## Generating class-specific localization maps ##########
CUDA_VISIBLE_DEVICES=7 python main.py --model deit_small_MCTformerV2_patch16_224 \
                --data-set VOC12MS \
                --scales 1.0 \
                --img-list voc12 \
                --data-path Datasets/VOC2012 \
                --resume MCTformer_results/voc/MCTformer_v2/checkpoint_best.pth \
                --gen_attention_maps \
                --attention-type fused \
                --layer-index 3 \
                --visualize-cls-attn \
                --patch-attn-refine \
                --attention-dir MCTformer_results/voc/MCTformer_v2/attn-patchrefine \
                --cam-npy-dir MCTformer_results/voc/MCTformer_v2/attn-patchrefine-npy \
                --out-crf MCTformer_results/voc/MCTformer_v2/attn-patchrefine-npy-crf \

######## Evaluating the generated class-specific localization maps ##########
python evaluation.py --list voc12/train_id.txt \
                    --gt_dir Datasets/VOC2012/SegmentationClassAug \
                    --logfile psa/voc12/evallog.txt \
                    --type npy \
                    --curve True \
                    --predict_dir MCTformer_results/voc/MCTformer_v2/attn-patchrefine-npy  \
                    --comment "train1464"
```

图 4. 一阶段参数配置

```
CUDA_VISIBLE_DEVICES=7 python psa/train_aff.py --weights res38_cls.pth \
                       --voc12_root Datasets/VOC2012 \
                       --la_crf_dir MCTformer_results/MCTformer_v2/attn-patchrefine-npy-crf_1 \
                       --ha_crf_dir MCTformer_results/MCTformer_v2/attn-patchrefine-npy-crf_12 \


CUDA_VISIBLE_DEVICES=7 python psa/infer_aff.py --weights resnet38_aff.pth \
                     --infer_list voc12/train_id.txt \
                     --cam_dir MCTformer_results/MCTformer_v2/attn-patchrefine-npy \
                     --voc12_root Datasets/VOC2012 \
                     --out_rw MCTformer_results/MCTformer_v2/pgt-psa-rw \

CUDA_VISIBLE_DEVICES=7 python evaluation.py --list voc12/train_id.txt \
                     --gt_dir Datasets/VOC2012/SegmentationClassAug \
                     --logfile MCTformer_results/MCTformer_v2/pgt-psa-rw/evallog.txt \
                     --type png \
                     --predict_dir MCTformer_results/MCTformer_v2/pgt-psa-rw \
                     --comment "train 1464"
```

图 5. 二阶段参数配置

```
CUDA_VISIBLE_DEVICES=7 python seg/train_seg.py --network resnet38_seg \
                       --num_epochs 30 \
                --seg_pgt_path /MCTformer_results/voc/MCTformer_v2/pgt-psa-rw  \
                       --init_weights res38_cls.pth  \
                       --save_path seg_log  \
                       --list_path /voc12/train_aug_id.txt \
                       --img_path /Datasets/VOCdevkit/VOC2012/JPEGImages  \
                       --num_classes 21 \
                       --batch_size 4 \

CUDA_VISIBLE_DEVICES=7 python seg/infer_seg.py --weight res38_cls.pth \
                        --network resnet38_seg  \
                        --list_path /voc12/val_id.txt  \
                        --gt_path /Datasets/VOCdevkit/VOC2012/SegmentationClassAug  \
                        --img_path /Datasets/VOCdevkit/VOC2012/JPEGImages  \
                        --save_path val_ms_crf \
                        --save_path_c val_ms_crf_c \
                        --scales 0.5 0.75 1.0 1.25 1.5 \
                        --use_crf True \
```

图 6. 三阶段参数配置

4.3 界面分析与使用说明

一阶段包括对 MCTformer-V2 模型的训练、生成 CAM 及对 CAM 的评估阶段（源代码缺失），二阶段包括对 ResNet-38 模型的训练、伪标签的生成及评估，三阶段包括 ResNet-38 分割模型的训练及分割推理。

部分实验运行过程如图 7和 8所示：

```
Epoch: [59]  [150/165]  eta: 0:00:11  lr: 0.000010  cls_loss: 0.2070 (0.2078)  attn_loss: 1.9619 (1.9621)  pat_loss
: 0.6934 (0.6934)  loss: 2.8629 (2.8633)  time: 0.6556  data: 0.0057  max mem: 13271
Epoch: [59]  [160/165]  eta: 0:00:03  lr: 0.000010  cls_loss: 0.1997 (0.2074)  attn_loss: 1.9547 (1.9613)  pat_loss
: 0.6934 (0.6934)  loss: 2.8535 (2.8621)  time: 0.5633  data: 0.0356  max mem: 13271
Epoch: [59]  [164/165]  eta: 0:00:00  lr: 0.000010  cls_loss: 0.1988 (0.2071)  attn_loss: 1.9547 (1.9608)  pat_loss
: 0.6934 (0.6934)  loss: 2.8478 (2.8613)  time: 0.5648  data: 0.0348  max mem: 13271
Epoch: [59] Total time: 0:02:01 (0.7350 s / it)
```

图 7. 一阶段训练过程

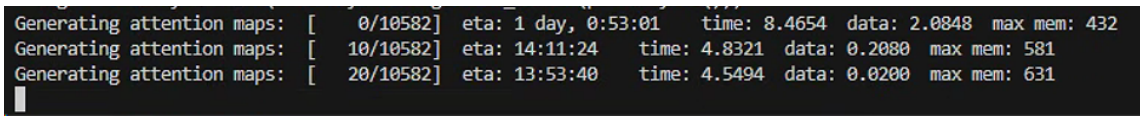

图 8. 一阶段生成 CAM 过程

5 实验结果分析

复现实验结果如图 9， 10和 11所示：

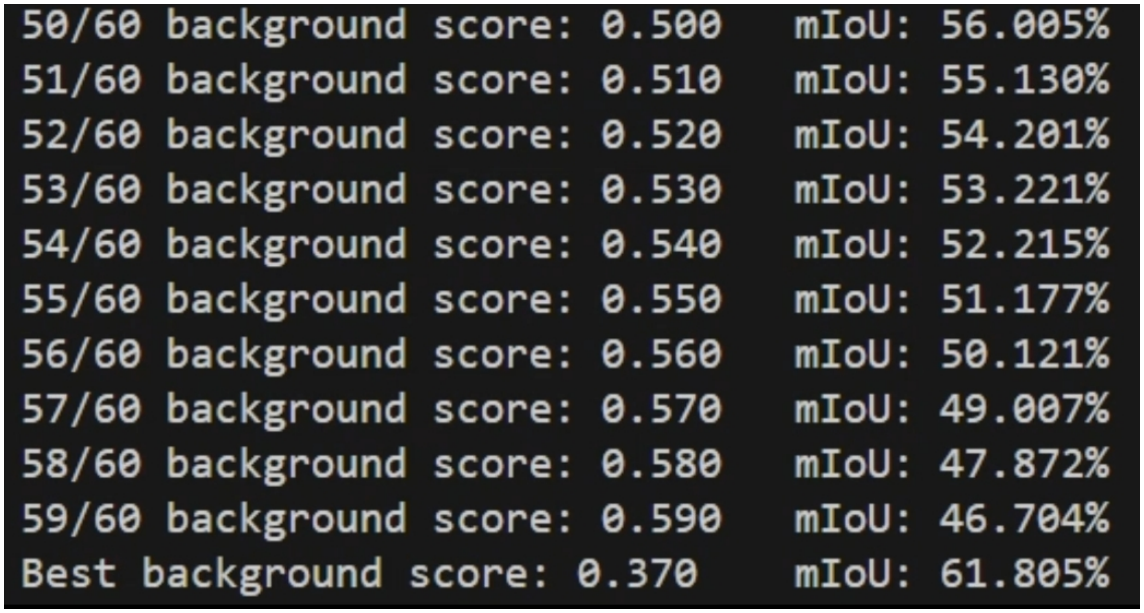

图 9. 一阶段评估结果图

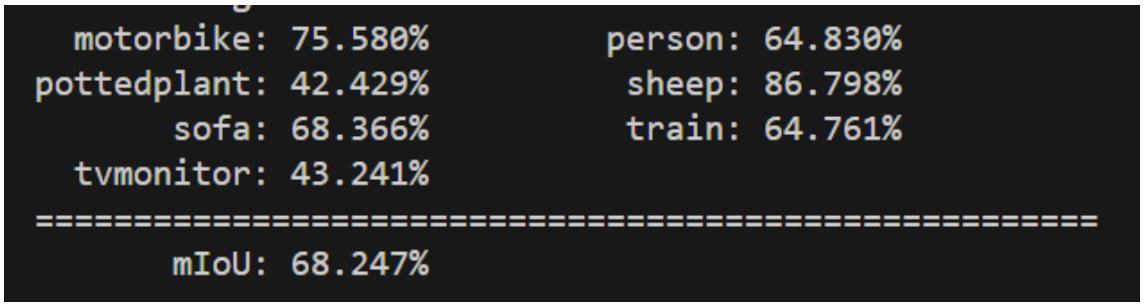

图 10. 二阶段评估结果图

图 11. 三阶段评估结果图

从图 9可以看出，论文中提供的结果为 61.7，一阶段复现结果达到 61.805，然而二阶段论文提供的结果为 69.1，复现达到 68.247，少于 1% 之内，三阶段的工作论文提供的结果为 71.9，复现结果为 69.0，少于接近 3%。

此外，方案一与方案二的实验结果分别如图 12和 13所示：

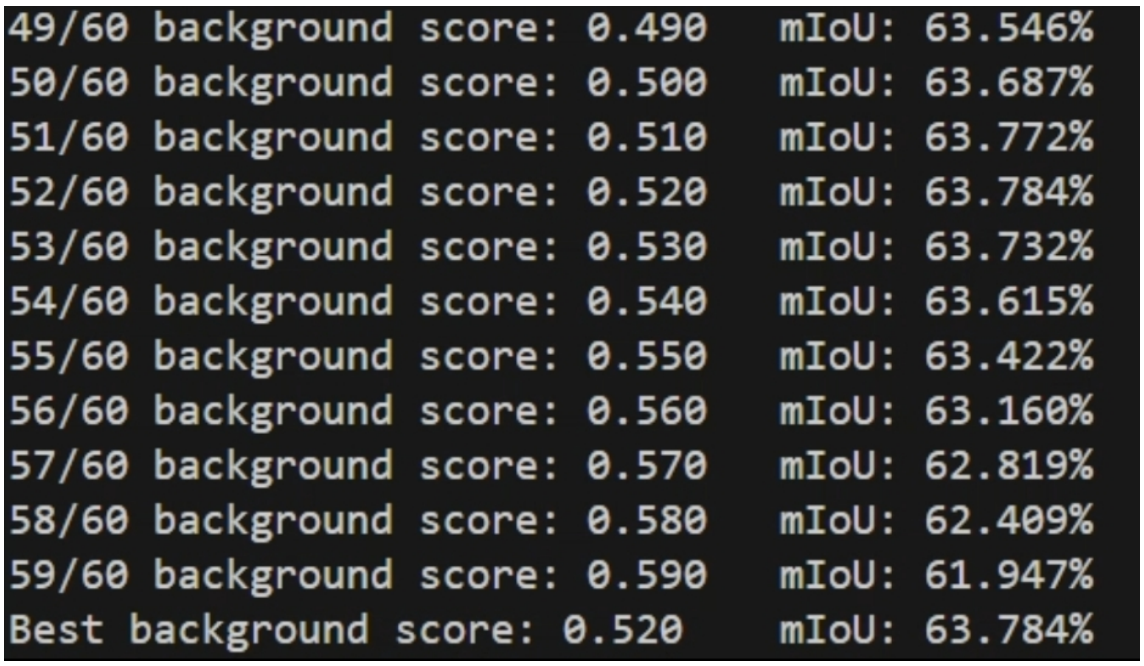

图 12. 方案一改进实验结果

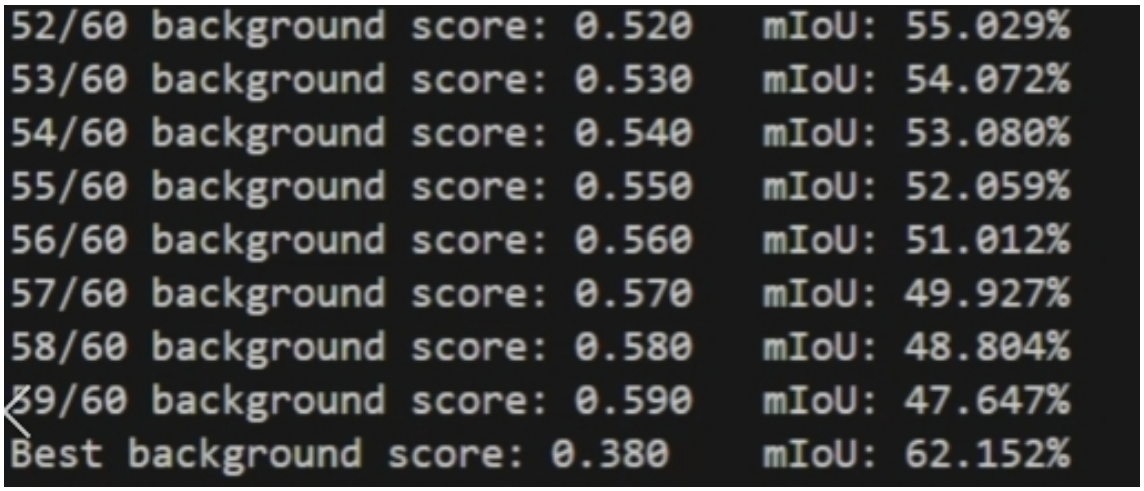

图 13. 方案二改进实验结果

与原论文提供的 61.7 的实验结果相比，方案一与方案二均有所提升，且方案一提升更为显著，这说明对于全局信息的提取能够进一步改善 MCTformer 模型，有助于模型找到非判别性区域。同时，转置的亲合度矩阵同样有助于细化生成的 CAM。

6 总结与展望

本文主要关注弱监督语义分割任务，提出了基于 Transformer 的改进方法 MCTformer-V1 和 MCTformer-V2。通过利用 Transformer 注意力和 CAM 模块，这些方法能够生成特定类别的目标定位图，并在训练过程中采用类感知训练策略和一致性增强策略来提高模型的学习效果。此外，本文还引入了 U-Net 网络来提取图像的全局特征，以减少判别性特征的丢失。通过实验验证，这篇论文的方法在弱监督语义分割任务中确实取得了较好的效果，具有一定的

研究价值和应用前景。此外，我所提出的两个改进方案也能够使得模型进一步得到提升，这也将成为我未来的研究工作之一。但不可否认，三阶段的复现工作比原论文低将近 3%，这里的复现工作仍需进一步改善，也会努力联系作者提供一些指导。

参考文献

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018.

[2] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8988–8997, 2020.

[3] Junsong Fan, Zhaoxiang Zhang, Tieniu Tan, Chunfeng Song, and Jun Xiao. Cian: Cross-image affinity net for weakly supervised semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:10762–10769, 04 2020.

[4] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X. Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. *ArXiv*, abs/2105.00957, 2021.

[5] Alexander Kolesnikov and Christoph Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. volume 9908, pages 695–711, 10 2016.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[7] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2020.

[8] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. 03 2018.

[9] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio' , and Yoshua Bengio. Graph attention networks. *ArXiv*, abs/1710.10903, 2017.

[10] Yude Wang, Jie Zhang, Meina Kan, S. Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12272–12281, 2020.

[11] Lian Xu, Wanli Ouyang, Bennamoun, Farid Boussaid, Ferdous Sohel, and Dan Xu. Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation.

2021 IEEE/CVF International Conference on Computer Vision (ICCV), pages 6964–6973, 2021.

[12] Bingfeng Zhang, Jimin Xiao, Yunchao Wei, Mingjie Sun, and Kaizhu Huang. Reliability does matter: An end-to-end weakly supervised semantic segmentation approach. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:12765–12772, 04 2020.

[13] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7222–7231, 2021.

[14] Tianyi Zhang, Guosheng Lin, Weide Liu, Jianfei Cai, and Alex Chichung Kot. Splitting vs. merging: Mining object regions with discrepancy and intersection loss for weakly supervised semantic segmentation. In *European Conference on Computer Vision*, 2020.

[15] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. 12 2016.

[16] Tianfei Zhou, Liulei Li, Xueyi Li, Chun-Mei Feng, Jianwu Li, and Ling Shao. Group-wise learning for weakly supervised semantic segmentation. *IEEE Transactions on Image Processing*, PP:1–1, 12 2021.