# CFR-ICL：用于交互式图像分割的级联前向细化与迭代点击损失

Shoukun Sun[1], Min Xian[1],Fei Xu[2],Luca Capriotti[2],和 Tiankai Yao[2]

[1]机器智能与数据分析实验室，计算机科学系，爱达荷大学[2]爱达荷国家实验室

{ssun,mxian}@uidaho.edu，{fei.xu,luca.capriotti,tiankai.yao}@inl.gov

arXiv:2303.05620v2 [cs.CV] 4 Mar 2024

## 摘要

基于点击的交互式分割旨在通过用户点击的引导，从图像中提取感兴趣的对象。近期研究通过利用输出结果提供的反馈，取得了出色的整体性能。然而，在大多数最先进方法中，1）推理阶段涉及缺乏灵活性的启发式规则，并且需要单独的细化模型；2）用户点击次数与模型性能无法兼顾。为应对这些挑战，我们提出了一种基于点击且由掩膜引导的交互式图像分割框架，其中包含三个创新组件：级联前向细化（CFR）、迭代点击损失（ICL）和 SUEM图像增强。CFR 提供了一个统一的推理框架，能够以由粗到细的方式生成分割结果。所提出的 ICL 使模型训练能够同时提升分割效果并减少用户交互。所提出的 SUEM 增强是一种全面的方法，可为交互式图像分割创建大规模且多样化的训练集。大量实验表明，该方法在五个公开数据集上达到了最先进的性能。值得注意的是，与此前最先进方法相比，我们的模型在伯克利和 DAVIS 数据集上，将达到 0.95 交并比所需的点击次数分别减少了 33.2% 和 15.5%。

## 1 引言

交互式图像分割通过用户输入从图像中提取感兴趣的对象。它对于推动基于深度学习的解决方案在现实场景中的广泛应用至关重要。深度神经网络（DNN）通常需要大量标注数据，而由于人工成本高昂，创建这些数据可能既昂贵又耗时。交互式分割为生成大规模标注数据集提供了一种更具成本效益的解决方案。

交互式分割中的交互方式包括涂鸦 [26],、框 [12], 和点击 [25]。本文聚焦于基于点击的方法，用户在图像上进行正点击和负点击，以分别标识前景和背景区域。早期的基于点击的方法采用图像处理技术开发，例如 [25] 中描述的基于连通性的方法。然而，随着该领域的后续发展，基于深度学习的方法应运而生，首先是 [28], ，它显著提升了分割性能。近年来，更多基于深度学习的方法相继问世，包括 [10, 23, 22, 1],，进一步提高了交互式分割的效率。

许多交互式分割方法 [10, 23, 1, 22, 16]根据预测分割掩码与真实标注之间的差异定义损失函数，却没有明确的方法来优化用户点击次数。一些研究[1] 通过细化初始预测来提升输出质量。这些技术需要额外的可训练网络模块，以识别并精细处理预测掩码中的局部区域，从而增加了训练难度。本文提出迭代点击损失（ICL），对点击次数较多的实例施加惩罚。我们提出一种在推理阶段进行细化的新策略，称为级联前向细化（CFR）。CFR 旨在推理过程中改善分割细节，无需额外的网络。此外，我们还提出一种有效的图像增强技术——SUEM 复制粘贴（C&P）。该技术是简单复制粘贴 [6] 方法的扩展

，

专为交互式图像分割任务设计。本文的贡献总结如下。

- 据我们所知，所提出的迭代点击损失是首个将点击次数编码在内、用于训练交互式分割模型的损失；它还为定义偏好用户点击次数较少的模型提供了一种新方法。

- 所提出的级联前向细化在推理过程中以一种简单统一的框架提升分割质量，并可应用于其他迭代式掩码引导交互式分割模型。

- 在所提出的 SUEM 增强中，我们提出了四种复制粘贴增强方法，可在实际场景中显著提升训练集的多样性并扩大其规模。

## 2 相关工作

**交互式分割**。在深度学习技术广泛应用之前，交互式分割主要通过

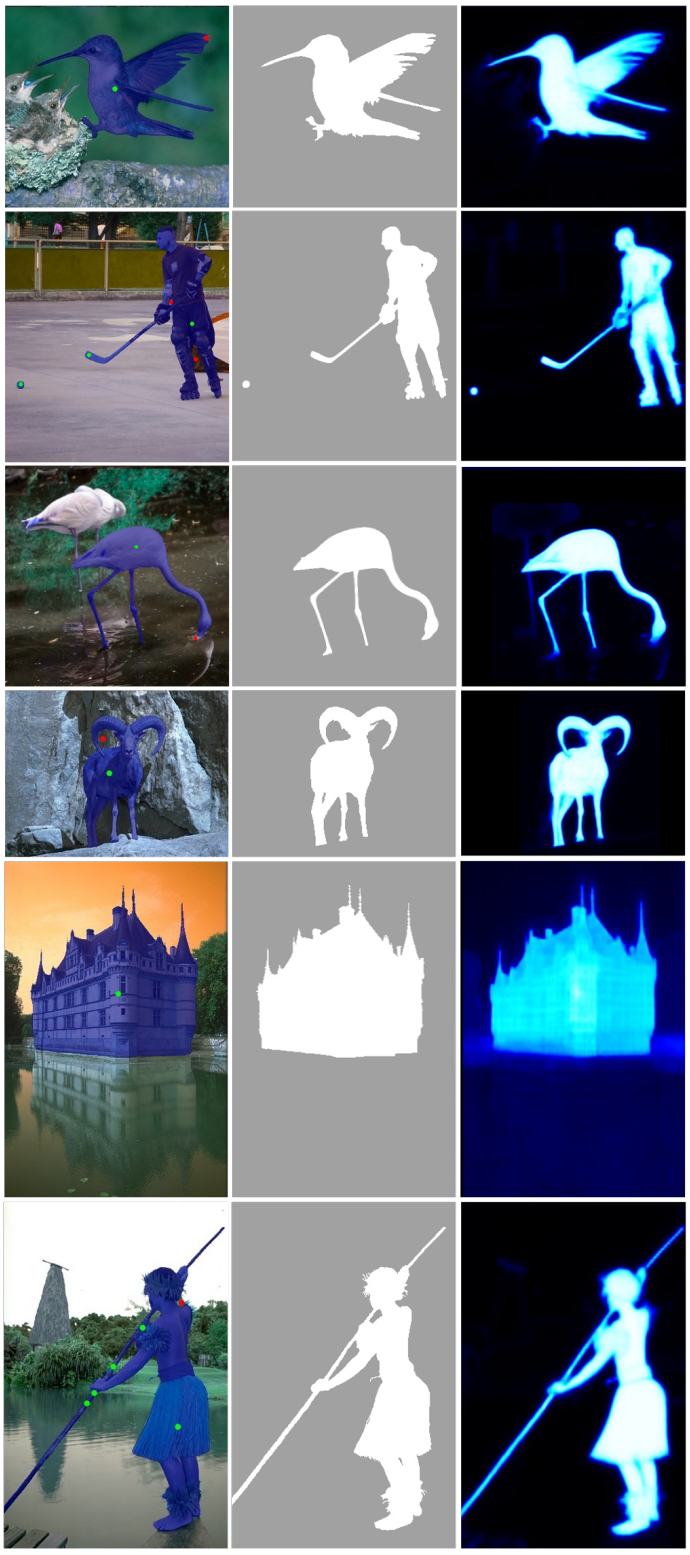


图 1：交并比超过 0.95 的分割结果示例。第一列展示带有点击标记的图像（绿色表示前景，红色表示背景），以及所提方法生成的分割掩码（蓝色）。第二列展示真实标注。第三列展示所提方法生成的概率图。这些原始图像来自伯克利[18] 和 DAVIS [19] 数据集。

基于图像处理的方法实现，例如 GrabCut [20]，NC-Cut [26]，和 EISeg [25]。随着深度学习的兴起，基于深度学习的交互式分割模型越来越受欢迎。DIOS 模型 [28] 将背景和前景用户点击编码为两张距离图，并将图像与这两张图拼接后作为输入。BRS [10] 和 f-BRS [23] 将该任务建模为推理时在线优化问题。BRS 在推理过程中优化图像或距离图，以改进给定用户点击后的分割结果；而 f-BRS 优化中间层权重，以实现相同目标并加快推理速度。FCA-Net [15] 构建了首击注意力模块，用于增强首次点击的重要性，因为它通常对确定主体位置起着关键作用。UCP-Net [4] 提出了一种新颖的基于轮廓的方法，要求用户在物体轮廓上进行点击。

**迭代式掩码引导交互分割**。[17] 提出了一种迭代采样策略，每次从误分类区域的中心采样一个点。这些迭代生成的点用于训练，以提升性能。RITM [22] 在训练过程中采用并改进了迭代采样策略，并迭代地将前一次输出作为模型输入，从而获得更高的分割质量。迭代式掩码引导模型仅涉及简单的前馈过程，相较于 BRS 和 f-BRS 等推理时优化方法，计算效率更高。近期，FocalClick [1] 提出了采用 SegFormer [27] 作为骨干网络的由粗到细流程，并取得了最先进的结果。该方法利用启发式策略确定可能包含错误的局部区域，并使用局部细化网络更新这些区域。SimpleClick [16] 采用经 MAE [9]，预训练的 Plain Vision Transformer（Plain ViT）作为 RITM 方法的骨干网络，大幅提升了性能。先进的 ViT网络架构显著促进了交互式分割的发展。我们的工作基于 RITM 和 SimpleClick 方法，进一步探究交互式分割的本质。

**一切分割模型**。Segment Anything Model (SAM) [11] 代表了图像分割领域的一项先进进展。SAM 建立在交互式分割的基础之上，支持多种用户交互方式。其设计以「可提示分割任务」为核心，能够根据文本、掩码、点击和框等多种提示生成有效的分割掩码。SAM 将大型图像编码器、灵活的提示编码器与快速的掩码解码器相结合，即使提示存在歧义，也能预测多个目标掩码。仅使用点和掩码作为输入时，SAM 有望发挥类似基于点击的交互式分割模型的作用。为评估其在基于点击的交互中的效果，我们将 SAM 模型与其他领先的基于点击的交互式分割模型进行比较。

**图像增强**。深度学习应用通常需要大量数据才能实现最佳模型性能，而训练样本的多样性至关重要。因此，开发高效的数据增强技术以提升数据利用效率十分必要。先前的研究 [6] 表明，简单复制粘贴策略可作为实例级分割任务的一种强大数据增强方法。我们针对交互式分割任务，开发了更具针对性的图像增强方法。

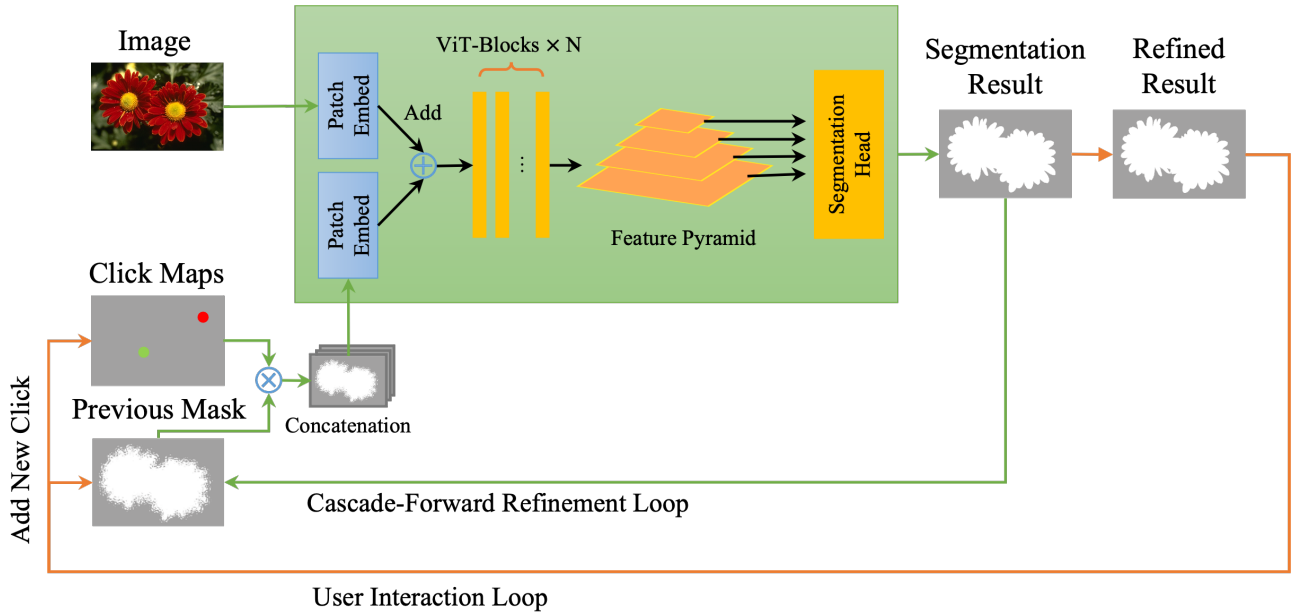



图 2：与级联前向细化相结合的迭代式掩码引导交互分割概览。橙色线条表示用户交互循环（外循环）。绿色线条表示细化循环（内循环）。黑色线条表示两个循环共享的过程。在用户交互循环中，用户会添加新的点击。在 CFR 循环中，利用点击对先前的掩码进行迭代优化。

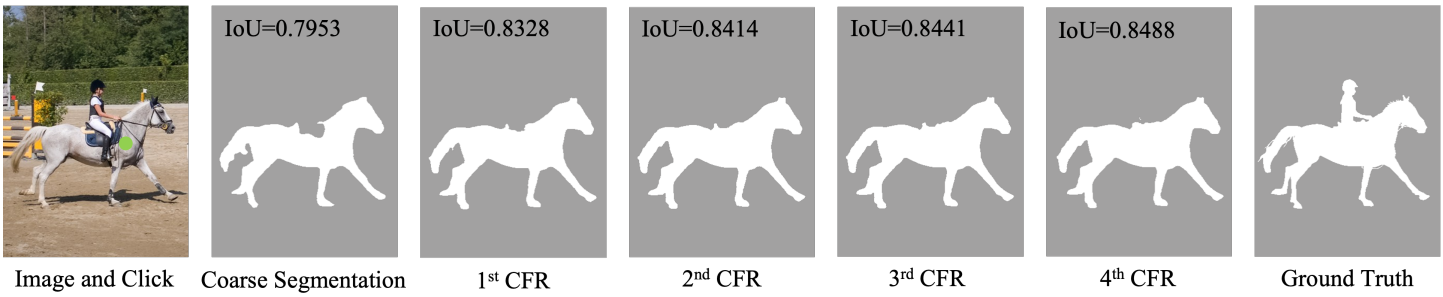



图 3：级联前向细化的示例结果。

## 3 所提方法

我们构建了一种新颖的迭代式掩码引导框架，其中包括：1）一种新颖的训练损失，将点击次数编码到损失项中，并使模型倾向于使用更少的点击；2）一种无需额外模块的推理时细化方案；3）一种针对交互式分割场景设计的有效图像增强方法。

### 3.1 迭代点击损失

Sofiiuk等人 [23] 将每张图像生成的所有用户点击编码为两个圆盘图，并在训练期间将其输入分割网络。其训练过程与传统图像分割深度神经网络的训练过程并无不同。在 [17, 22], 中，研究人员采用了一种混合策略，将随机采样的点击与迭代生成的点击结合起来，以生成输入图。不过，训练得到的模型 1）在推理期间可能需要更多用户交互；2）缺乏有效方法来平衡输入点击次数与分割性能。

为克服这些挑战，我们提出迭代点击损失（ICL）方法，将点击次数融入训练过程。首先，根据真实标注生成一组随机采样的初始点击，记为$P^0$，将其输入模型以获得初始输出 $Y^0$。随后，从$Y^0$中的误分类区域进行采样，逐次生成点击（每次一个），并将新生成的点击与 $P^0$ 结合，形成新的序列 $P^1$。重复这一过程，生成完整的点击序列 $P^t$。点击采样策略可表示为

$$
\begin{aligned}
P^0 &= \text{Random sample from } \mathbb{Y} \\
Y^0 &= f(X, P^0, \mathbf{0}) \\
P^1 &= P^0 \cup S(\mathbb{Y}, Y^0) \\
Y^1 &= f(X, P^1, Y^0) \\
&\dots \\
P^t &= P^{t-1} \cup S(\mathbb{Y}, Y^{t-1}) \\
Y^t &= f(X, P^t, Y^{t-1}),
\end{aligned}
\tag{1}
$$

其中，$\mathbf{0}$ 是零初始化的掩码，$\mathbb{Y}$ 是真实标注，并且




(a) Source and extra (b) Simple (c) Union (d) Exclusion (e) Image Mixing

图4：复制粘贴模式示意图。

$S$ 是一个采样函数 [22]，用于在输出掩码中的误分类区域内选取一个点击点。传统的总损失函数 [22] 定义为

$$L = \mathbb{L}(Y^t, \mathbb{Y}). \tag{2}$$

其中，$\mathbb{L}$ 为归一化焦点损失 [21]。请注意，在公式 2 中，使用一系列点击集合 $[P^0, P^1, ..., P^t]$ 和一系列输出 $[Y^0, Y^1, ..., Y^{t-1}]$ 来生成分割掩码 $Y^t$；然而，只有最终输出用于计算归一化焦点损失并更新模型参数。在提出的 ICL 方法中，我们构建了一种新的损失函数，用于累积生成的掩码序列的加权损失。

$$L_{ICL} = \sum_{i=1}^{t} \beta_i \mathbb{L}(Y^i, \mathbb{Y}), \tag{3}$$

其中，$\beta_i$ 用于控制每一项的权重。上述公式中的每次点击都会产生一个损失项，最小化损失既能提升分割性能，也能同时减少点击次数。通过提高点击次数较多时对应损失项的权重（较大的 $i$），模型会倾向于使用更少的点击来实现精准分割。ICL 为交互式分割模型提供了一种新颖的方法，使其能够优先考虑减少用户点击次数。

## 3.2 级联前向细化

已有类似的由粗到细的流程 [1]，在对整幅图像进行初始粗分割后，通过引入局部细化模块来改善局部区域的细节。这些由粗到细的策略整体表现出色；然而，这些方法 1）依赖启发式规则来选择需要进一步细化的图像区域；2）必须分别独立训练两个深度学习模型。

我们提出了一种基于级联前向细化（CFR）的推理策略，无需使用两个模型即可迭代细化分割结果。所提出的 CFR 包含两个推理循环：外循环通过逐步增加用户交互生成粗分割掩码，内循环则使用相同的输入图像和用户点击，多次前向传递分割模型以细化掩码。

令 $f$ 表示用于分割的深度神经网络，$Y^t$ 表示第 $t$ 步的输出，$X$ 表示输入图像。在第 $t$ 步的用户点击序列记为 $P^t = \{(u_k, v_k, l_k)\}_{k=1}^{t}$，其中 $(u_k, v_k)$ 表示用户点击的坐标，$l_k \in \{0, 1\}$ 表示其标签（0 表示背景，1 表示前景）。模型的输入包括原始图像、由点击生成的图，以及先前的分割掩码。我们将 CFR 定义为

$$Y_0^t = f(X, P^t, Y_n^{t-1}), \text{and} \tag{4}$$

$$Y_i^t = f(X, P^t, Y_{i-1}^t), i \in \{1, 2, 3, ..., n\}, \tag{5}$$

其中，$Y_0^t$ 表示第 $t$ 个粗分割步骤（外循环）的输出，而 $Y_n^{t-1}$ 表示步骤 $t-1$ 时最后得到的细化掩码。公式 5 定义了内循环中第 $i$ 步的细化结果，$n$ 表示细化步骤的数量，该数量可以设为固定值，也可以自适应确定。级联前向细化方法利用公式 4 和 5 迭代更新粗分割掩码和细化掩码。在细化过程中（内循环）无需额外的用户点击，分割掩码会持续得到细化，从而提供更高质量的输入。图 2 展示了 CFR 流程的概览，图 3 展示了一系列 CFR 细化结果。

**固定步长 CFR 和自适应 CFR 推理**。固定步长 CFR 对每次用户点击执行 $n$ 次细化。推理过程中会应用此细化，以提升输出质量，并且无需修改模型即可集成到任何迭代式掩码模型中。除了固定步长方法，我们还提出并验证了一种自适应 CFR（A-CFR）方案。该方案会统计 $Y_i^t$ 与 $Y_{i-1}^t$ 之间发生变化的像素数量；当变化像素数低于指定阈值或达到最大步数时，便终止内循环。我们分别用 CFR-$n$ 和 A-CFR-$n$ 表示固定 $n$ 步 CFR 和最大步数为 $n$ 的自适应 CFR。

## 3.3 SUEM 复制粘贴

为生成适用于交互式分割场景的大规模、多样化数据集，我们提出了一种全面的图像增强方法，即 SUEM 复制粘贴（C&P），该方法包含四种 C&P 模式：简单复制粘贴 [6]，并集复制粘贴、排除复制粘贴和图像混合。

C&P 方法的基本原理是将从一张图像中随机选取的对象插入另一张图像中。在交互式分割中，我们将感兴趣的对象及其对应的图像分别称为源对象 和

| | GrabCut | | 伯克利 | | DAVIS | | Pascal VOC | | SBD | |
|---|---|---|---|---|---|---|---|---|---|---|
| NoC@ | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 |
| 模型 | | | | | | | | | | |
| SimpleClick | 1.54 | 2.16 | 2.46 | 6.71 | 5.48 | 12.23 | 2.81 | 3.75 | 5.24 | 11.23 |
| ICL | 1.50 | 2.00 | 2.34 | 6.48 | 5.44 | 11.86 | **2.62** | **3.58** | **5.05** | **11.03** |
| SUEM C&P | **1.44** | **1.74** | **1.97** | **5.08** | **5.28** | **10.69** | 2.68 | 3.59 | 5.33 | 11.48 |

表1：ICL 和 SUEM C&P 的有效性。

源图像。相反，从训练集中随机选取的对象及其对应图像分别称为额外对象 和额外图像。

**简单复制粘贴模式**。 简单复制粘贴模式会将源对象插入随机选取的额外图像中，并以源对象的掩码作为真实标注。图4（b）直观展示了此模式。

**并集复制粘贴模式**。 交互式分割通常需要识别由多个对象组成的目标对象，例如一个男人拥抱一个孩子。为模拟此类情形，我们采用并集复制粘贴模式，将额外对象粘贴到源图像中，并将它们各自掩码的并集作为真实标注。图4（c）展示了此模式生成的图像。

**排除复制粘贴模式**。 另一种常见情形是，感兴趣的对象被另一个对象遮挡，例如一个人站在柱子后面。为解决这一问题，我们引入排除复制粘贴模式，将额外对象复制到源图像中，并以源对象的掩码作为真实标注，同时排除额外对象的掩码。图4（d）直观展示了此模式。

**图像混合模式**。 图像混合方法将源图像与另一张图像混合，并将源对象的掩码用作真实标注。图像混合模式如图4（e）所示

上述策略结合起来，用于生成我们实验中的训练图像。

| | RITM | EMC-Click | Ours |
|---|---|---|---|
| GrabCut | 2.48 | **1.94** | 1.96 |
| 伯克利 | 5.41 | 6.92 | **5.05** |
| DAVIS | 11.52 | 11.26 | **10.85** |
| Pascal VOC | - | **3.44** | 3.57 |
| SBD | 12.00 | **11.68** | 12.02 |

表 2：RITM、EMC-Click [3] 和所提出方法的 NoC@95。采用 HRNet32 骨干网络和训练集 C+L。

# 4 实验

## 4.1 数据集与实验设置

**数据集**。以下图像集用于评估我们方法的性能。

- **GrabCut [20]:** 该数据集包含 50 张图像，每张图像中均有一个实例。
- **伯克利 [18]:** 该数据集包含 96 张图像，涵盖 100 个实例。
- **DAVIS [19]:** 该数据集包含从[10]50 个视频中提取的 345 张图像，并附有高质量的分割掩码。
- **PascalVOC [5]:** 仅使用验证集，包含1449张图像和3427个实例。
- **SBD [8]:** 训练集包含8497张图像和20172个实例。验证集包含2857张图像和6671个实例。
- **COCO [14]+LVIS [7] (C+L)：** COCO数据集的训练子集中包含99k张图像和120万个实例。LVIS数据集包含100k张图像和总计120万个实例。我们从COCO和LVIS合成了一个组合数据集C+L [22] ，包含104k张图像和160万个实例，仅用于训练。

**评估指标**。我们使用点击次数（NoC）指标评估模型性能。这一标准指标量化了用户为获得令人满意的分割结果所需进行的输入操作，且这些结果必须超过预先设定的交并比（IoU）阈值。本研究报告NoC@90和NoC@95。尽管早期研究通常报告NoC@85和NoC@90阈值，但对高质量分割日益增长的需求使得模型评估标准更加严格。我们采用 [13] 中所述的方法在评估过程中生成点击。点击次数最多限制为20次。

**骨干网络模型**。按照 SimpleClick [16], 提出的方法，我们在实验中采用了 SimpleClick ViT-Base 和 ViT-Huge（ ViT-B 和 ViT-H）模型。这些模型使用 Plain Vision Transformer [2] 作为骨干网络。随后，我们利用所提方法对这些预训练模型进行了微调。为进一步探索我们的方法的适应性，我们还将实验扩展至采用 HRNet [24] 骨干网络。

**实现细节**。我们的迭代点击学习（ICL）使用迭代生成的三个点击进行模型训练，设置为 $\beta_i \in [1, 2, 3]$。我们采用Adam优化器，参数为 $\beta_1 = 0.9$, $\beta_2 = 0.999$，学习率为 $5 \times 10^{-6}$。所有模型均在训练好的 SimpleClick 模型基础上微调一个 epoch。ViT-B 的批量大小设为 140，ViT-H 的批量大小设为 32。我们采用归一化焦点损失（NFL） [21] 进行训练，参数设为 $\alpha = 0.5, \gamma = 2$。除了提出的 SUEM C&P 增强方法外，我们还使用了包括缩放、翻转、旋转、亮度和对比度调整以及裁剪在内的图像增强方法。输入图像统一调整为 $448 \times 448$。点击被编码为半径为 5 的圆盘图。所有模型均在 NVIDIA Quadro RTX 8000 GPU 上训练。

| | GrabCut | | 伯克利 | | DAVIS | | Pascal VOC | | SBD | |
|---|---|---|---|---|---|---|---|---|---|---|
| NoC@ | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 |
| 推理 | | | | | | | | | | |
| StdInfer | 1.54 | 2.16 | 2.46 | 6.71 | 5.48 | 12.23 | 2.81 | 3.75 | 5.24 | 11.23 |
| CFR-1 | **1.44** | **2.04** | **2.35** | **6.43** | 5.33 | 11.99 | **2.70** | **3.58** | **5.11** | **11.14** |
| CFR-4 | 1.50 | 2.18 | 2.38 | 6.46 | 5.31 | 11.94 | 2.72 | 3.59 | 5.16 | 11.16 |
| A-CFR-4 | 1.50 | 2.06 | 2.39 | **6.43** | **5.28** | **11.90** | 2.73 | 3.60 | 5.17 | **11.14** |

表 3：CFR方案的结果。“StdInfer”表示标准推理 [22]；CFR-$n$ 表示步长固定为 $n$的 CFR；A-CFR-$n$表示最大步数为 $n$的自适应 CFR。

| | GrabCut | | PascalVOC | | SBD | | 伯克利 | | DAVIS | |
|---|---|---|---|---|---|---|---|---|---|---|
| NoC@ | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 | 90 | 95 |
| 模型 | | | | | | | | | | |
| !DIOS（全卷积网络） | 6.04 | - | 6.88 | - | - | - | 8.65 | - | - | - |
| !FCA-Net-SIS（Res2Net） | 2.08 | - | 2.69 | - | - | - | 3.92 | - | 7.57 | - |
| #LD（VGG-19） | 4.79 | - | - | - | 10.78 | - | - | - | 9.57 | - |
| #BRS（DenseNet） | 3.60 | - | - | - | 9.78 | - | 5.08 | - | 8.24 | - |
| #f-BRS-B（ResNet101） | 2.60 | 4.82 | - | - | 7.05 | 13.81 | 4.13 | 10.05 | 7.31 | 14.30 |
| #RITM（HRNet18） | 2.04 | 3.66 | - | - | 5.42 | 11.65 | 3.23 | 8.38 | 6.70 | 13.88 |
| #CDNet（ResNet34） | 2.64 | - | - | - | 7.87 | - | 3.69 | - | 6.66 | - |
| #UCP-Net（EffNet） | 2.76 | - | - | - | - | - | 2.70 | - | - | - |
| #PseudoClick（HRNet18） | 2.04 | - | 2.74 | - | 5.40 | - | 3.23 | - | 6.57 | - |
| #FocalClick (HRNet18) | 2.06 | - | - | - | 6.52 | - | 3.14 | - | 6.48 | - |
| #FocalClick (SegF-B0S2) | 1.90 | - | - | - | 6.51 | - | 3.14 | - | 7.06 | - |
| #SimpleClick（ViT-H $^{\text{Baseline}}$） | 1.44 | 2.10 | 2.20 | 3.01 | 4.15 | 9.86 | 2.09 | 6.34 | 5.33 | 11.67 |
| #SimpleClick CFR-1（ViT-H） | **1.32** | 1.78 | **2.16** | 2.92 | **4.08** | **9.80** | 2.15 | 6.21 | 5.22 | 11.64 |
| #我们的方法 (ViT-H) | 1.46 | 1.66 | 2.20 | 3.00 | 4.45 | 10.52 | 1.77 | 4.73 | 4.86 | 9.06 |
| #我们的方法 CFR-1 (ViT-H) | 1.42 | **1.62** | 2.17 | **2.91** | 4.45 | 10.50 | **1.74** -16.7% | **4.44** -30.0% | **4.77** -10.5% | **8.85** -24.2% |
| -SAM（ViT-H） | 1.84 | 2.18 | 2.72 | 3.71 | 7.59 | 14.97 | 2.06 | 4.64 | 5.21 | 9.87 |
| *RITM（HRNet18） | 1.54 | 2.22 | - | - | 6.05 | 12.47 | 2.26 | 6.46 | 5.74 | 12.45 |
| *RITM (HRNet32) | 1.56 | 2.48 | - | - | 5.71 | 12.00 | 2.10 | 5.41 | 5.34 | 11.52 |
| *PseudoClick (HRNet32) | 1.50 | - | 2.25 | - | 5.54 | - | 2.08 | - | 5.11 | - |
| *FocalClick（SegF-B3S2） | 1.52 | 1.84 | 2.89 | 3.80 | 5.63 | 11.58 | 1.93 | 4.55 | 4.96 | 10.71 |
| *SimpleClick（ViT-H $^{\text{Baseline}}$） | 1.50 | **1.66** | 1.98 | 2.51 | 4.70 | 10.76 | 1.75 | 4.34 | 4.78 | 8.88 |
| *SimpleClick CFR-1（ViT-H） | 1.56 | 1.76 | **1.94** | 2.46 | **4.60** | **10.74** | 1.67 | 4.20 | 4.72 | 8.76 |
| *我们的方法 (ViT-H) | **1.48** | **1.66** | 1.99 | 2.52 | 4.81 | 10.94 | 1.51 | **2.90** | 4.27 | 7.62 |
| *我们的方法 CFR-1 (ViT-H) | 1.58 | 1.76 | **1.94** | **2.45** | 4.74 | 10.90 | **1.46** -16.6% | **2.90** -33.2% | **4.24** -11.3% | **7.50** -15.6% |

表4：与最先进方法的比较。’!’表示在 Pascal VOC [5] 数据集上训练的模型。’#’表示在 SBD [8] 数据集上训练的模型。’-’表示在 SA-1B [11] 数据集上训练的模型，’*’表示在 C+L[22] 数据集上训练的模型。’-’表示该值不可用。以上标形式显示的百分比表示相较于对应基线模型的点击次数减少比例。

## 4.2 ICL 和 SUEMC&P 的有效性

我们在 SBD 数据集上，使用提出的 ICL 和 SUEM C&P 方法对最初在 SBD 数据集上训练的 SimpleClick ViT-Base 模型进行了微调。表 1 对比了原始 SimpleClick 模型与通过 ICL 和 SUEM C&P 微调后的模型的性能。如表 1 所示，ICL 提升了 SimpleClick 模型在全部五个数据集上的表现。相比之下，SUEM C&P 显著提升了模型在 GrabCut、伯克利和 DAVIS 数据集上的效果，凸显了 ICL 和 SUEM C&P 的优势。

为进一步探究所提训练方法的适用性，我们还将 ICL+SUEM C&P 应用于 HRNet32，并将其与 RITM 和 EMC-Click 进行比较。表 2 展示了使用 ICL+SUEM C&P 和 HRNet32 骨干网络得到的结果。这些结果表明，与 RITM 模型相比，在 GrabCut、伯克利和 DAVIS 数据集上采用所提方法时，NoC@95 分别显著降低了 20.97%、6.65% 和 5.82%。所提方法在 GrabCut、Pascal VOC 和 SBD 上取得了有竞争力的结果，并在伯克利和 DAVIS 数据集上显著优于 EMC-Click。这表明 ICL 和 SUEM C&P 与不同骨干网络具有良好的兼容性。

## 4.3 使用 CFR方案进行推理

表 3 验证了 CFR 推理过程的有效性。我们在 SBD 测试集上，对使用 SBD 训练的 SimpleClick ViT-B 模型进行推理时，采用了固定步长 CFR 和自适应 CFR（A-CFR）。如表 3 所示，三种 CFR 推理方案（CFR-1、CFR-4 和 A-CFR-4）均优于标准推理方法（“StdInfer”） [22], , 由此证明 CFR 能够提升交互式分割性能。

尽管 CFR-4 消耗了更多计算资源，但相较于 CFR-1 的提升并不明显，这表明增加步数并不一定能带来更好的结果。A-CFR-4 会在变化低于设定阈值时自适应地停止细化；本例中的阈值为 20 像素。CFR-1 和 A-CFR-4 的性能相当，但在效率和适应性方面有所不同。例如，CFR-1 仅执行一个步骤，而 A-CFR-4 可以通过调整阈值来适应特定情况。评估计算效率后，我们在后续实验中选择了 CFR-1。

### 4.4 与最先进方法的比较

表4详细展示了我们提出的模型与十二种先进的基于深度学习的交互式分割方法的性能。我们的方法（“我们的方法”）结合 ICL 和 SUEM C&P，对 ViT-H 模型进行微调。在表4中，以双实线分隔，并用符号“!”、“#”、“-”和“*”标识的模型，分别是在 Pascal VOC、SBD、SA-1B [11], 和 C+L 数据集上训练的模型。由于 Pascal VOC 数据集较小，研究中很少报告使用该数据集进行训练的结果。

表4显示，在 SBD 和 C+L 上训练的 SimpleClick ViT-H 模型，在五个测试数据集上的 NoC@90 和 NoC@95 指标均优于先前方法。因此，我们选取这些模型作为基线，记为“基线”。

对于在 SBD 上训练的模型，将我们提出的 CFR-1 推理方法应用于 SimpleClick 基线模型后，除伯克利数据集上的 NoC@90 外，所有指标均显著下降。具体而言，在 GrabCut 上，CFR-1 使 NoC@90 和 NoC@95 分别降低 8.3% 和 15.2%。采用 ICL 和 SUEM C&P 构建的所提模型“我们的方法 CFR-1”，将达到交并比 0.9 所需的点击次数（NoC@90）减少了 16.7%。值得注意的是，与基线相比，NoC@95 降低了 30.0%，表明该方法在高精度图像分割方面有效。不过，在 GrabCut、PascalVOC 和 SBD 测试集上并未取得显著提升。这可能是因为 GrabCut 和 PascalVOC 中的视觉内容较为简单，而 SBD 的真实标注不够准确，因此在高交并比阈值下难以显著减少用户点击次数。

在 C+L 数据集上训练的模型也观察到了类似的结果。SimpleClick CFR-1 模型在除 GrabCut 外的所有数据集上均提升了 NoC@90 和 NoC@95。‘我们的方法 CFR-1’模型在 Berkeley 和 DAVIS 数据集上的提升尤为显著。

## 5 结论

本研究提出了一种基于深度神经网络（DNN）的交互式图像分割框架，包含三个创新组件：1）基于迭代点击损失（ICL）的训练；2）基于级联前向细化（CFR）的推理；3）SUEM 复制粘贴图像增强。该框架也适用于其他迭代式掩码引导交互分割方法。所提出的 ICL 引入了一种减少点击次数的创新方法。CFR 推理采用统一框架，通过两个循环迭代细化分割结果。所提出的 SUEM C&P 是一种全面的图像增强方法，可生成更多样化的训练集。我们进行了全面实验以验证所有组件的有效性，结果表明，所提出的方法达到了最先进的性能。

## 参考文献


[1] Xi Chen 等人。“FocalClick：迈向实用的交互式图像分割”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2022 年，第 1300–1309 页。[2] Alexey Dosovitskiy 等人。“一张图像胜过 16×16 个词：用于大规模图像识别的 Transformer”。载于：国际学习表征会议。2021 年。网址：https : // openreview . net/ forum ? id =YicbFdNTTy。[3] Fei Du 等人。“面向基于点击的交互式图像分割的高效掩码校正”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2023 年 6 月，第 22773–22782 页。[4] Camille Dupont、Yanis Ouakrim 和 Quoc Cuong Pham。“UCP-Net：用于实例分割的非结构化轮廓点”。载于：*2021 IEEE*系统、人类与控制论国际会议（*SMC*）。2021 IEEE系统、人类与控制论国际会议（SMC）。2021 年，第 3373–3379 页。DOI：10.1109/SMC52423

.

2021.9658754。

[5] Mark Everingham 等人：《Pascal 视觉对象类别（VOC）挑战》。载于：国际计算机视觉期刊88.2（2010），第 303–338 页。

[6] Golnaz Ghiasi 等人：《简单复制粘贴是一种强大的实例分割数据增强方法》。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2021，第 2918–2928 页。

[7] Agrim Gupta、Piotr Dollar 和 Ross Girshick：《LVIS：大词汇量实例分割数据集》。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2019，第 5356–5364 页。

[8] Bharath Hariharan 等人：《来自逆检测器的语义轮廓》。载于：*IEEE/CVF* 国际计算机视觉大会论文集。IEEE，2011，第 991–998 页。DOI：10 . 1109 / ICCV . 2011 . 6126343。

[9] 何恺明等人：《掩码自编码器是可扩展的视觉学习器》。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2022 年，第 16000–16009 页。

[10] Won-Dong Jang 和 Chang-Su Kim。“通过反向传播细化方案实现交互式图像分割”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2019年，第5297–5306页。[11] Alexander Kirillov等人。“分割一切”。载于：*IEEE/CVF*国际计算机视觉大会论文集。2023年10月，第4015–4026页。[12] Victor Lempitsky等人。“利用边界框先验进行图像分割”。载于：*IEEE/CVF*国际计算机视觉大会论文集。2009年IEEE第12届国际计算机视觉会议。2009年，第277–284页。DOI：10.1109/ICCV.2009. 5459262。

[13] Zhuwen Li、Qifeng Chen 和 Vladlen Koltun。“具有潜在多样性的交互式图像分割”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2018年，第577–585页。[14] Tsung-Yi Lin等人。“Microsoft COCO：上下文中的常见物体”。载于：欧洲计算机视觉会议。Springer，2014年，第740–755页。[15] Zheng Lin等人。“基于首次点击注意力的交互式图像分割”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2020年，第13339–13348页。[16] Qin Liu等人。“SimpleClick：基于简单视觉变换器的交互式图像分割”。载于：*IEEE/CVF*国际计算机视觉大会论文集。2023年10月，第22290–22300页。[17] Sabarinath Mahadevan、Paul Voigtlaender 和 Bastian Leibe。“迭代训练的交互式分割”。载于：英国机器视觉会议（*BMVC*）。2018年。[18] D. Martin等人。“人工分割自然图像数据库及其在评估分割算法和测量生态统计中的应用”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。第八届IEEE国际计算机视觉大会论文集。ICCV 2001。第2卷。2001年，第416–423页，第2卷。DOI：10.1109/ICCV. 2001.937655。[19] F. Perazzi等人。“视频目标分割基准数据集与评估方法”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2016年IEEE计算机视觉与模式识别会议（CVPR）。2016年，第724–732页。DOI：10.1109/CVPR.2016.85。[20] Carsten Rother、Vladimir Kolmogorov 和 Andrew Blake。“GrabCut”：使用迭代图割进行交互式前景提取。载于：*ACM*图形学汇刊23.3（2004年），第309–314页。ISSN：0730-0301。DOI：10 . 1145 / 1015706. 1015720。URL：https : / / doi . org/10.1145/1015706.1015720（访问日期：2022-03-09）。[21] Konstantin Sofiiuk、Olga Barinova 和 Anton Konushin。“AdaptIS：自适应实例选择网络”。载于：*IEEE/CVF*国际计算机视觉大会论文集。2019年，第7355–7363页。[22] Konstantin Sofiiuk、Ilya A. Petrov 和 Anton Konushin。“利用掩码引导重振交互式分割的迭代训练”。载于：*IEEE*国际图像处理会议论文集。2022年IEEE国际图像处理会议（ICIP）。2022年，第3141–3145页。DOI：10.1109/ICIP46576. 2022.9897365。[23] Konstantin Sofiiuk等人。“F-Brs：重新思考交互式分割中的反向传播细化”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2020年，第8623–8632页。[24] Ke Sun等人。“用于人体姿态估计的深度高分辨率表征学习”。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2019年6月，第5686–5696页。DOI：10. 1109/CVPR.2019.00584。[25] Min Xian等人。“EISeg：高效交互式分割”。载于：国际模式识别会议（*ICPR*）。2016年第23届国际模式识别会议（ICPR）。2016年，第1982–1987页。DOI：10. 1109/ICPR.2016.7899927。[26] Min Xian 等人：《中性连通性切割》。载于：*IEEE* 图像处理汇刊25.10（2016），第 4691–4703 页。ISSN：1941-0042。DOI：10.1109/TIP.2016. 2594485。[27] Enze Xie 等人：《SegFormer：用于语义分割的变换器简洁高效设计》。载于：神经信息处理系统进展34（2021），第 12077–12090 页。[28] Ning Xu 等人：《深度交互式目标选择》。载于：*IEEE/CVF* 计算机视觉与模式识别会议论文集。2016，第 373–381 页。DOI：10.1109/CVPR.2016.47。