南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Abductive multi-instance multi-label learning for periodontal disease classification with prior domain knowledge

Zi-Yuan Wu, Wei Guo, Wei Zhou, Han-Jia Ye, Yuan Jiang, Houxuan Li, Zhi-Hua Zhou

Medical Image Analysis, 2025

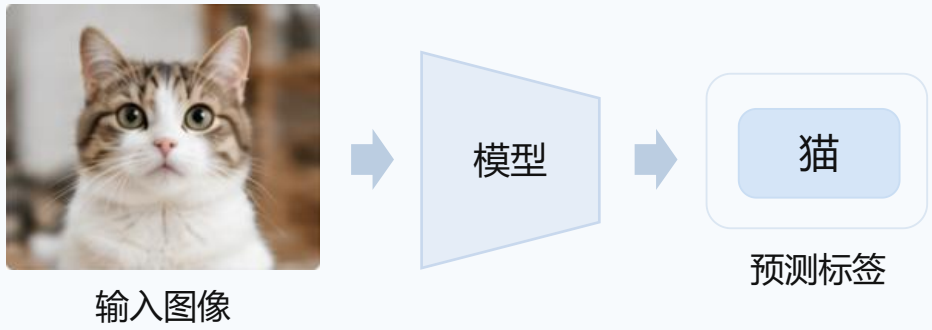
汇报人：王必广

汇报日期：2026.09.22

Multi-Instance Multi-Label Learning (MIML)

Traditional Supervised Learning

一个样本对应一个标签



一个输入 → 一个标签

Multi-Instance Multi-Label Learning (MIML)

一个样本包含多个实例，并对应多个标签

一个输入（多个实例） → 多个标签

MIML 的核心思想： 用于处理 **“一个样本由多个实例组成，同时对应多个标签”** 的学习任务，能够建模输入中的局部区域与多个标签之间的复杂对应关系。

Two Major Paradigms in AI

人工智能中的两大范式

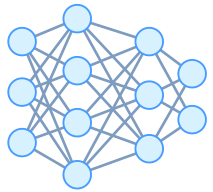
Connectionism（连接主义）

输入数据（例如：图像）



猫　狗　鸟 …

神经网络



预测结果

猫	**0.90**
狗	0.06
鸟	0.04
…	

从数据中学习输入到输出的映射

- 数据驱动
- 感知能力强
- 但缺少显式逻辑推理

Symbolism（象征主义）

知识表示

苏格拉底 → 是人
人 → 会死亡
发烧 → 是一种症状
咳嗽 → 是一种症状
…

规则 / 逻辑

如果 X 是人，
则 X 会死亡。

如果 发烧 ∧ 咳嗽，
则 可能肺炎。

…

推理结果

苏格拉底会死亡

诊断：可能肺炎

…

- 知识驱动
- 推理能力强
- 但难以直接处理原始图像

连接主义：擅长“看”

从数据中学习，具有强大的感知能力

＋

象征主义：擅长“想”

基于知识和规则，具有强大的推理能力

»

人工智能的发展趋势：从两大范式走向融合

Neuro-symbolic AI = Connectionism + Symbolism

越来越多的研究开始尝试将神经网络的感知能力与符号规则的推理能力结合起来。

Abductive Learning (ABL)

不仅关注最终结果，还约束中间解释过程

传统机器学习

输入图片

模型 → 最终预测：猫

只监督最终结果是否正确

结果对了，就认为预测正确

溯因学习（ABL）

1 预测阶段

模型输出最终结果 + 预定义中间概念标签

输入图片

模型 → 最终预测：猫

预定义概念空间

has_ear	= 1
has_whisker	= 1
has_fur	= 1
color_yellow	= 1
has_tail	= 0

2 知识校验阶段

利用规则检查中间标签是否合理

知识库 / 规则

cat -> has_ear ∧ has_whisker ∧ has_tail

若预测为猫，则应具有耳朵、胡须、尾巴等特征

溯因推理

修正后的概念标签

原始： has_tail = 0 ✗

修正： has_tail = 1 ✓

z_pred = [1, 1, 1, 1, 0]

z_abd = [1, 1, 1, 1, 1]

3 反馈学习阶段

将修正后的标签作为监督信号

Loss（基于修正标签）

反向传播

更新模型

利用修正后的标签进行学习

ABL 的核心思想： 模型在预定义的概念空间中进行预测，知识库对中间概念标签进行校验与修正，再将修正后的标签作为监督信号反向训练模型。

本质上，ABL = 机器学习 + 知识库 + 溯因推理

研究背景：牙周疾病患病负担持续上升

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Research Background

口腔疾病影响范围广，牙周炎已成为值得重点关注的公共口腔健康问题。

35亿

全球口腔疾病影响人数接近 35 亿

WHO Global Oral Health Status Report, 2022

约40%

30岁以上人群中，约 40% 受到牙周炎影响

Eke et al., 2016

- 牙周疾病是常见且具有长期危害的口腔疾病之一
- 早期筛查与规范诊断对口腔健康管理具有重要意义
- 这为智能化、标准化辅助诊断方法的研究提供了现实需求

数据来源：论文引言引用 WHO (2022) 与 Eke et al. (2016) 相关统计结果。

研究任务：基于口腔图像的牙周疾病分类诊断

1 传统临床诊断依据

1. **病史询问**
   了解患者病史
2. **临床检查**
   评估牙周组织状况
3. **牙周袋深度测量**
   牙周探诊等临床测量
4. **牙槽骨丧失评估**
   结合检查结果综合判断

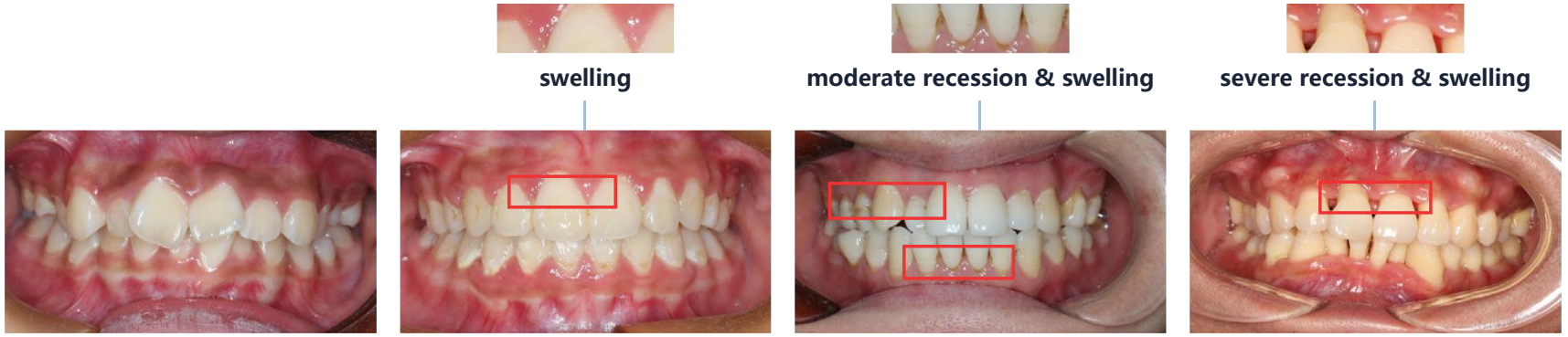
原文指出的局限

部分检查具有侵入性；
测量结果可能因检查者及
所用探针不同而存在差异。

本文的研究任务

利用口腔图像进行牙周疾病分类，
探索辅助诊断的标准化方法。
图像分类不等同于完整临床诊断。

2 本文关注的四类牙周疾病



Periodontal Health (PH) | Gingivitis (GI) | Mild to Moderate Periodontitis (MMP) | Severe Periodontitis (SP)

本文按肿胀与退缩属性组合，对四类标签进行建模。

类别	疾病名称	典型临床表现
PH	牙周健康 (Periodontal Health)	牙龈状态正常，无明显肿胀或牙周退缩
GI	牙龈炎 (Gingivitis)	存在牙龈肿胀（swelling），但无明显牙周退缩
MMP	轻中度牙周炎 (Mild to Moderate Periodontitis)	存在牙龈肿胀，并伴随中度牙周退缩 (moderate recession)
SP	重度牙周炎 (Severe Periodontitis)	存在牙龈肿胀，并伴随严重牙周退缩 (severe recession)

现有方法：基于目标检测的牙周疾病诊断

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

1 基本思路

1. **原始口腔图像**

2. **目标检测模型**
   RCNN / Faster R-CNN / YOLO

3. **定位疑似病灶区域**
   Region of Interest (ROI)

4. **提取局部病灶特征**

5. **输出整图疾病类别**
   PH / GI / MMP / SP

局部定位 → 特征提取 → 整图分类

2 两阶段诊断流程

先定位局部病灶，再判断整图类别

A 输入口腔图像

原始口腔图像
(包含多颗牙齿及牙龈组织)

整图输入

B 检测病变区域

目标检测模型定位
疑似病变区域（ROI）

局部病灶检测

C 综合分类结果

- **PH** 牙周健康 Periodontal Health
- **GI** 牙龈炎 Gingivitis ✓
- **MMP** 轻中度牙周炎 Mild to Moderate Periodontitis
- **SP** 重度牙周炎 Severe Periodontitis

最终诊断
GI（牙龈炎）

现有方法的局限性：基于目标检测方法的不足

南京航空航天大学 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

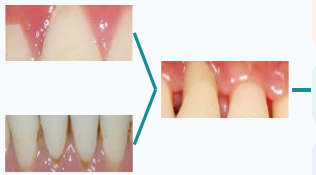
作者指出：现有检测式方法虽有效，但在牙周疾病诊断任务中仍存在明显局限。

检测式方法
先定位病灶 → 再判断类别

1 依赖精细标注，成本高

- 需要医生逐张标注病灶区域（ROI / bounding box）
- 细粒度标注耗时、费力，难以大规模获取
- 与本文仅有整图粗粒度标签的场景不匹配

2 难以刻画复杂病变关系



牙龈炎

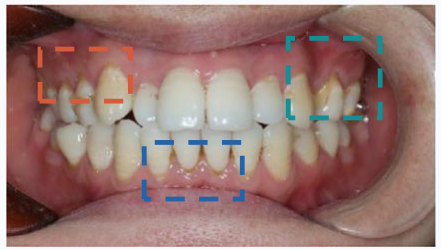
轻中度牙周炎

重度牙周炎

多个局部表现共同决定类别

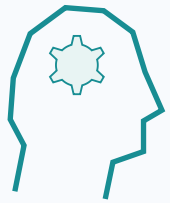
- 并非“单一区域 → 单一类别”
- 一个类别往往由多个局部表现共同决定
- 肿胀、退缩与严重退缩之间存在组合关系

3 忽略邻近区域与整体结构信息



- 诊断不仅看单个病灶，还要结合邻近区域状态
- 需要综合考虑多处可疑区域及其空间关系
- 仅检测局部框，难以反映完整诊断过程

4 难以融入医生先验知识



- 病变常见于特定区域
- 病变类型与位置相关
- 疾病具有严重程度层级
- .....

- 病灶通常出现在特定局部区域
- 医生具有病变位置和类别关系的经验知识
- 现有方法难以将这些先验知识显式纳入训练

因此，作者希望寻找一种新方法：既能建模“图像—局部区域—疾病标签”之间的复杂关系，又能利用医生的先验知识。

作者方法：AB-MIML 的核心设计思想

南京航空航天大学 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

1 类别重构

从疾病类别到子标签组合

疾病类别（原标签）

- PH
- GI
- MMP
- SP

临床属性（子标签）

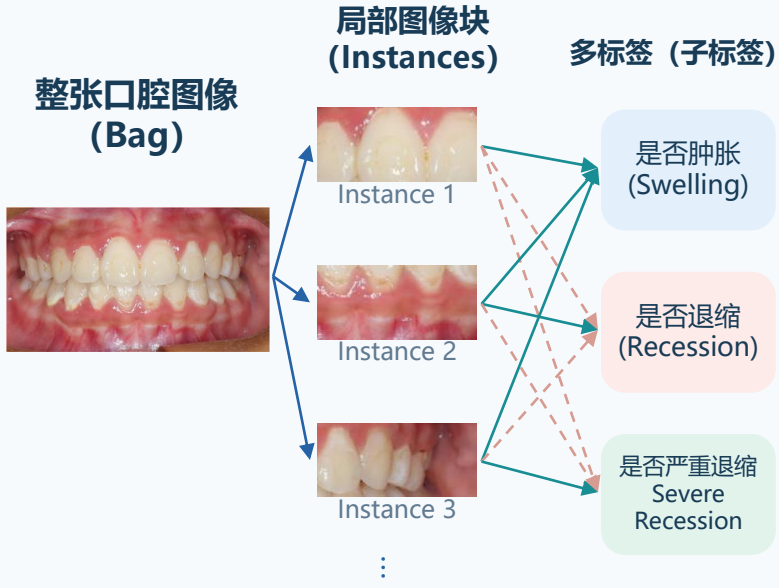
- 是否肿胀 Swelling
- 是否退缩 Recession
- 是否严重退缩 Severe Recession

疾病类别	肿胀	退缩	严重退缩
PH	0	0	0
GI	1	0	0
MMP	1	1	0
SP	1	0	1

疾病类别并非彼此完全独立，
而是可由临床属性组合描述。

2 MIML 建模

整图作为 Bag，局部 Patch 作为 Instance



建模图像—局部区域—疾病标签之间的关系，
采用多实例多标签学习（MIML）框架。

3 专家知识融入

逻辑公式 + 溯因推理

医生经验 图像结构信息 → 逻辑公式 → 知识库 KB → 溯因推理

示例规则（简化形式）

1. 健康图像 → 不应存在肿胀 / 退缩 patch。
2. 中心区域更可能出现病变。

基于知识的标签校正

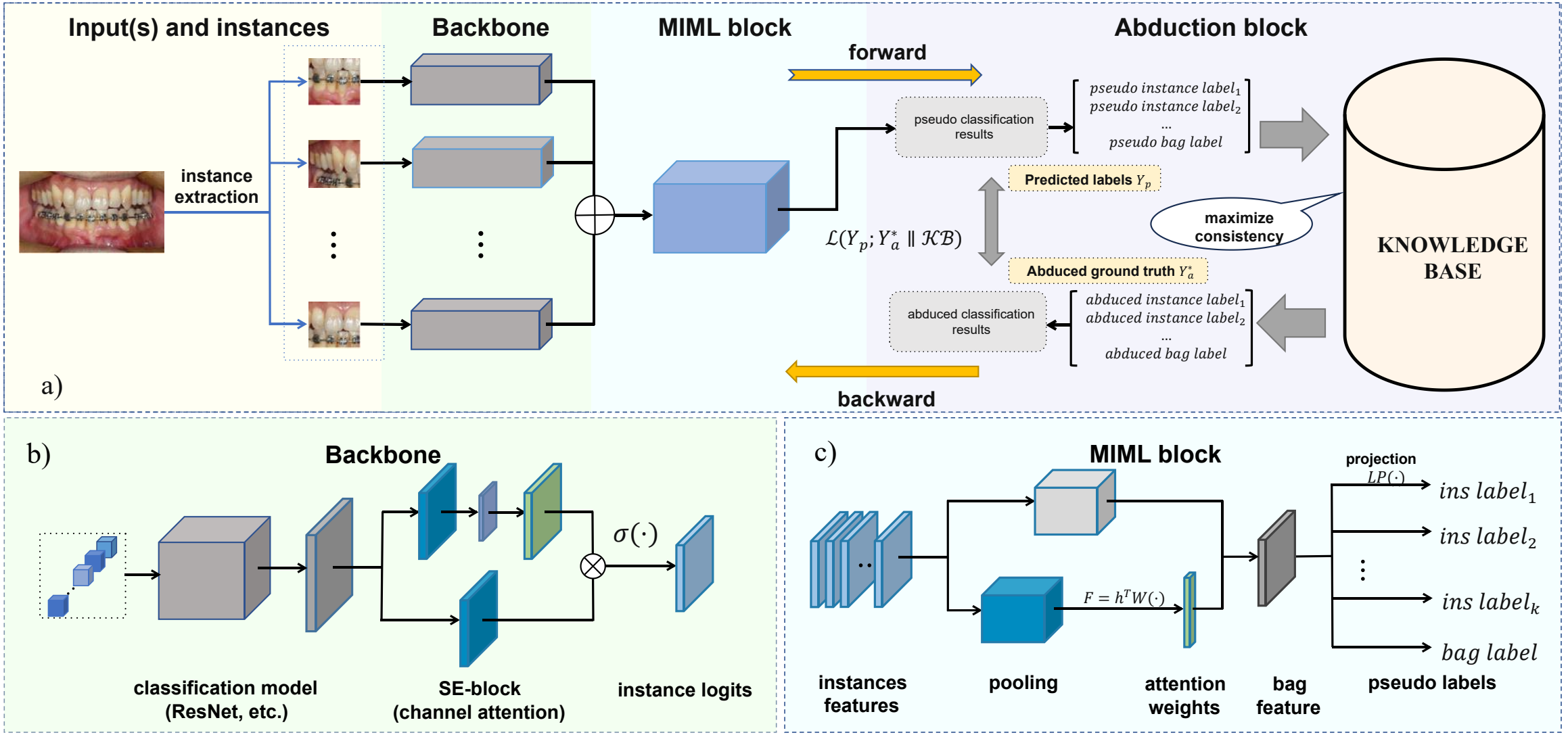
Pseudo labels → 与知识库一致性校正 → Abduced labels

利用先验知识修正预测结果，
提高可解释性与准确性。

AB-MIML = 类别分解 + 多实例多标签学习 + 先验知识推理

目标：同时建模图像区域、临床表现与疾病类别之间的关系。

AB-MIML 整体框架


Input(s) and instances
Backbone
MIML block
Abduction block
instance extraction
forward
pseudo classification results
pseudo instance label$_1$
pseudo instance label$_2$
...
pseudo bag label
Predicted labels Y_p
$\mathcal{L}(Y_p; Y_a^* \| \mathcal{KB})$
maximize consistency
KNOWLEDGE BASE
Abduced ground truth Y_a^*
abduced classification results
abduced instance label$_1$
abduced instance label$_2$
...
abduced bag label
backward
a)
b)
Backbone
$\sigma(\cdot)$
classification model (ResNet, etc.)
SE-block (channel attention)
instance logits
c)
MIML block
projection $LP(\cdot)$
$ins\ label_1$
$ins\ label_2$
$ins\ label_k$
$bag\ label$
$F = h^T W(\cdot)$
instances features
pooling
attention weights
bag feature
pseudo labels

问题建模

Problem Formulation

数据集定义

为进行常见牙周疾病的分类诊断，构建数据集：

$$D = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$$

- $X_i \subseteq \mathcal{X}$：第 i 个样本的口腔内图像（intraoral images）。
- $Y_i \subseteq \mathcal{Y} \in [C] = \{1, \ldots, C\}$：与图像 X_i 对应的疾病类别标签。
- 在牙周疾病诊断任务中，$C = 4$，对应四类疾病：牙周健康（PH）、牙龈炎（GI）、轻中度牙周炎（MMP）、重度牙周炎（SP）。
- 目标是从 D 中学习一个映射 $f: \mathcal{X} \to \mathcal{Y}$，能够对来自同一分布的新口腔图像做出准确的预测。

问题动机

- 与自然图像不同，牙周疾病的诊断依赖于图像中**多个特定的局部区域**，多个区域的诊断结果共同决定最终的疾病类别及其严重程度。
- 不同疾病类别之间**具有层次关系**，不能简单地视为相互独立的平行类别。
- 传统的“一张图像对应一个标签”的映射函数，**难以刻画**图像中**局部区域与疾病标签之间复杂的对应关系**。
- 因此，本文将牙周疾病的诊断任务建模为**多实例多标签学习（MIML）**，并进一步引入**专家先验知识**，通过**溯因学习（ABL）**对模型预测的伪标签进行推理与修正，以提高模型的准确性和可解释性。

AB-MIML 方法定义

AB-MIML Formal Definition

方法定义

- 1. 给定知识库 KB 与数据集

$$D = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)\}$$

- 2. 每个图像 X_i 被看作一个 bag，由多个局部实例组成：

$$X_i = \{x_{i1}, x_{i2}, \ldots, x_{i,n_i}\}$$

- 3. 每个标签 Y_i 为分解后的子标签集合：

$$Y_i = \{y_{i1}, y_{i2}, \ldots, y_{i,l}\}$$

- 4. 学习映射函数：

$$f: \mathcal{X} \to \mathcal{Y}$$

- 5. 满足约束：

$$\mathrm{KB} \models O(f)$$

- 6. 其中 $O(f)$ 表示由模型预测 grounded 的事实，含义是预测结果应与知识库保持一致。

Backbone: 分类模型与通道注意力

模块说明

1. **输入局部图像块（patch）**

   将整张图像划分为多个局部区域，每个 patch 作为一个 instance 输入分类模型。

2. **特征提取（分类模型）**

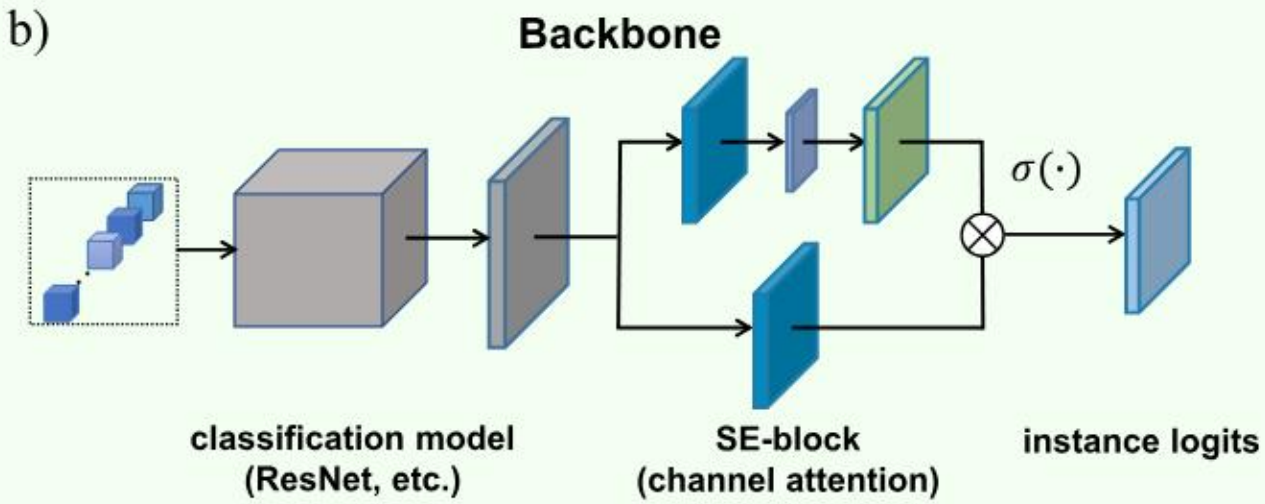
   采用预训练的分类网络（如 ResNet、ResNeXt、ViT 等），将每个 patch 映射为高维视觉特征。

3. **通道注意力增强（SE-block）**

   在分类模型后加入 SE-block，学习不同 feature channel 的权重，增强判别性特征，抑制无关信息。

4. **输出 instance-level 表示**

   得到增强后的特征表示（instance logits/feature），用于后续 MIML 模块进行 bag-level 与 instance-level 标签预测。



通道注意力公式

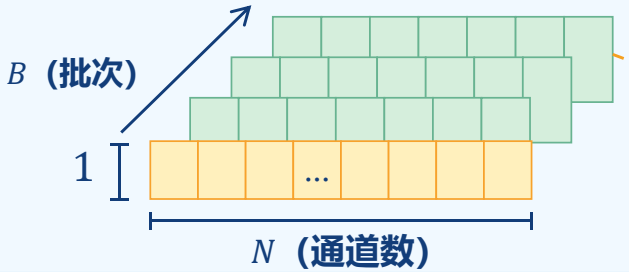
$$g_c = \tanh(f_c * W_S) * W_E$$

$$h_c = f_c \cdot g_c$$

Channel Attention: 公式计算过程示意

1 从张量到单个向量的理解

输入特征 f_c
$(B\times1\times N)$



为便于理解，从 batch 中取出一条向量

示例：一个 instance feature（$1\times N$）

$f_c = [0.2,0.8,0.1,0.6]$

0.2	0.8	0.1	0.6

2 Channel Attention 的计算过程（以一个向量为例）

输入向量

$f_c = [0.2,0.8,0.1,0.6]$

0.2	0.8	0.1	0.6

分支 A：生成通道权重（学习每个 channel 的重要性）

$g_c = \tanh(f_c * W_S) * W_E$

压缩映射 W_S $N \to d$ → tanh → 恢复映射 W_E $d \to N$

$z = [0.6, -0.2]$　　$t = [0.54, -0.20]$

$g_c = [0.9,0.2,0.7,0.1]$

0.9	0.2	0.7	0.1

分支 B：保留原始特征

$f_c = [0.2,0.8,0.1,0.6]$

逐元素相乘 ×

增强后的特征

$h_c = f_c \cdot g_c$

$h_c = [0.18,0.16,0.07,0.06]$

0.18	0.16	0.07	0.06

3 思考与解释

Q1 为什么 g_c 不能直接随机初始化一个与 f_c 同维的向量，而要由 f_c 计算得到？

A　因为注意力权重应随当前输入特征自适应变化；
不同 patch 的重要通道不同，因此 g_c 必须依赖当前 f_c 动态生成，而不是固定参数。

Q2 为什么不用直接 $g_c = f_c * W$，而要先压缩 → tanh → 恢复？

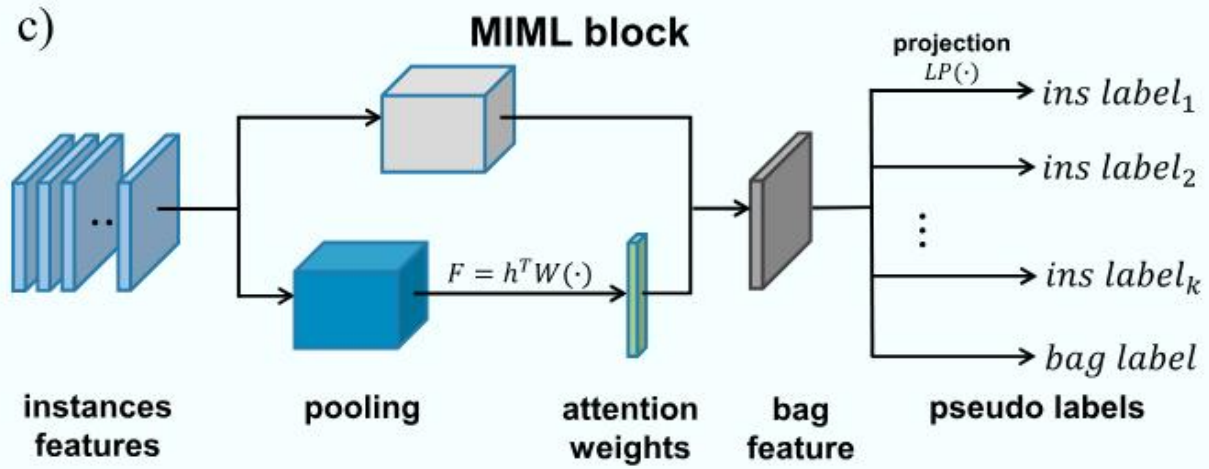
A　先降维可减少参数、提炼关键信息；tanh 引入非线性；
再升维可学习更灵活的通道依赖关系，比直接线性映射表达能力更强。

MIML Block: 模块作用与输入输出

模块作用

- 建模多实例输入与多标签输出之间的对应关系
- 对来自 Backbone 的多个 patch 特征进行注意力聚合
- 为后续伪标签预测提供 bag-level 表示

MIML block 结构



输入与输出

- 输入：多实例特征 f_m，形状为 $B \times K \times N$
- 其中 K 为 patch / instance 数，N 为特征维度
- 输出：聚合后的 bag feature h_m，形状为 $B \times 1 \times N$
- h_m 再送入线性层，生成 instance / bag 的 pseudo labels

核心理解

- 先学习每个 patch 的注意力权重 g_m
- 再对所有 instance feature 加权汇聚
- 得到 bag-level 表示后生成 pseudo labels

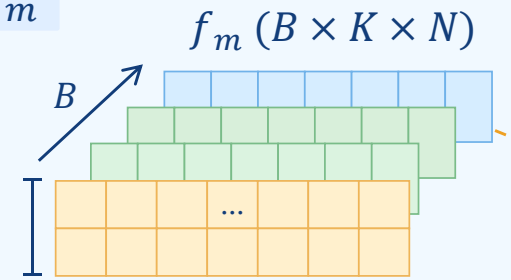
MIML Block: 注意力聚合与伪标签生成

1 输入：多实例特征 f_m

Backbone 输出
每张图像的多实例特征

$$f_m \in \mathbb{R}^{B\times K\times N}$$

B：批大小
K：实例数（patch 数）
N：特征维度



为便于理解，
仅分析一个样本

单个样本的实例特征 $f_m \in \mathbb{R}^{K\times N}$ **(例如** $K=4, N=4$**)**

f_1	0.2	0.8	0.1	0.6
f_2	0.5	0.3	0.9	0.2
f_3	0.6	0.2	0.4	0.7
f_4	0.1	0.7	0.5	0.3

K：实例数
（patch 数）
N：特征维度

2 注意力聚合（Attention MIL）

$$g_m^T = \mathrm{Softmax}(\sigma(f_m * W_{att}))\qquad h_m = g_m * f_m$$

核心思想：为每个 patch 分配重要性权重，
再对所有 instance feature 加权汇聚，得到 bag-level feature。

实例特征
$f_m \in \mathbb{R}^{K\times N}$

f_1 ... ; f_2 ... ; f_3 ... ; f_4 ...

$K \times N$

→

线性投影
$f_m * W_{att}$

W_{att} $(N\times 1)$

$K \times 1$

→

得分向量
$s = f_m * W_{att}$

0.3
1.2
2.1
0.8

$K \times 1$

→

非线性激活
$\sigma(\cdot) = \tanh(\cdot)$

0.29
0.83
0.97
0.66

$K \times 1$

→

归一化
$\mathrm{Softmax}(\cdot)$

0.10
0.20
0.50
0.20

$g_m^T \in \mathbb{R}^{K\times 1}$
(或 $g_m \in \mathbb{R}^{1\times K}$)

→

加权汇聚
$h_m = g_m * f_m$

$h_m = 0.10f_1 + 0.20f_2 + 0.50f_3 + 0.20f_4$

...

$1 \times N$

3 输出：伪标签生成

聚合后的 bag 特征
$h_m \in \mathbb{R}^{1\times N}$ (或 $B\times 1\times N$)

... $1 \times N$

→

多线性层
(共享输入嵌入)
Multiple Linear Layers /
Shared Embedding

→ ins label 1
→ ins label 2
...
→ ins label k
→ bag label

重加权特征 h_m 将被送入多个线性层，
生成 pseudo labels，并传递到
abduction block。

一句话理解

1 输入：K 个 instance feature
2 聚合：学习 patch 重要性并生成 bag feature
3 输出：预测 instance / bag 的 pseudo labels

Abduction Block: 基本思想与两大核心组件

核心思想

- 感知模块先从原始图像中预测 pseudo labels；
- 若预测结果与领域知识不一致，则利用逻辑溯因进行修正；
- 修正后的 abduced labels 再反馈给模型训练，从而提升预测性能。

形式化定义

$$\{x_1, \ldots, x_m\},\ f \triangleright \mathcal{O}$$

$$\text{s.t.} \qquad KB \models \mathcal{O}, \text{or}$$

$$KB \models \Delta(\mathcal{O}), f \leftarrow \psi(f, \Delta(\mathcal{O}))$$

- $\mathcal{O}$：模型预测得到的逻辑事实
- KB：知识库中的领域规则
- $\Delta(\mathcal{O})$：对原预测事实的修正

Abduction block 的两大核心组件

1. Knowledge Base

- 将口腔图像结构信息与医生专业知识转为逻辑公式
- 包含 bag–instance 对应规则与疾病诊断规则
- 论文中约包含 17 条逻辑公式

2. Abduction Process

- 将 pseudo labels 对齐为知识库中的逻辑变量
- 检查与知识库的一致性，并在代价约束下修正标签
- 输出 abduced labels，返回感知模块参与训练

Abduction Block: Knowledge Base 领域知识库

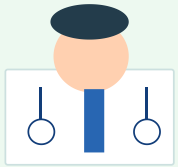
1 Knowledge Base 构建

我们将口腔图像的结构信息和医生的专业诊断知识转化为逻辑公式，构建领域知识库（Knowledge Base），作为溯因推理的先验知识。

口腔图像结构信息

- 牙齿的空间位置关系
- 邻域结构（如中心区域）
- 不同区域的解剖约束

+



医生专业诊断知识

- 疾病的发生规律
- 疾病之间的关联关系
- 经验性的诊断准则

领域知识库

(Knowledge Base)

- 约 17 条逻辑公式
- 覆盖主要的结构信息和牙周病诊断规则

2 知识库规则示例（Fig. 3）

知识库主要包含两类规则：bag-instance 对应关系规则和专业诊断知识规则。

1 Bag-instance correspondence（包–实例对应规则）

约束整体（bag）标签与局部（instance）标签之间的一致性。

例如：若患者整体为健康，所有实例级标签均不应包含疾病因素。

示例公式（论文 Fig. 3）：

$$\neg swelling(X) \wedge \neg recession(X) \leftarrow is\_bag(X) \wedge is\_healthy(X)$$
$$\neg swelling(Y) \wedge \neg recession(Y) \leftarrow instance\_of(X, Y) \wedge is\_healthy(X)$$

X：一个 bag（整张图像）
Y：一个 instance（图像块）
$is\_healthy(X)$：患者整体健康
$swelling(\cdot)$：牙龈肿胀
$recession(\cdot)$：牙龈退缩
$instance\_of(X,Y)$：Y 属于 X

2 Professional diagnosis rules（专业诊断知识规则）

利用医学领域知识，考虑疾病的空间关系和相互关联性。

例如：如果中心区域存在牙龈退缩，则相邻区域也可能存在退缩的高风险。

示例公式（论文 Fig. 3）：

$$\neg recession(Z) \leftarrow center(Y, Z) \wedge \neg recession(Y)$$
$$recession(Y) \leftarrow center(Y, Z) \wedge center(Y, Z') \wedge recession(Z) \wedge recession(Z')$$

Y, Z, Z'：不同的 instance
$center(Y,Z)$：Z 位于 Y 的中心区域
$recession(\cdot)$：牙龈退缩

Knowledge Base 将医生的领域经验转化为逻辑约束，为后续的溯因推理提供可靠的先验知识。

Abduction Process（溯因推理过程）

基于知识库约束，对模型预测的伪标签进行一致性检查，并在有限修改成本下生成更合理的标签。

1. 模型预测（伪标签）

对整张图像和局部 patch 进行预测，得到初始逻辑变量取值（可能包含错误）。

输入图像
论文 Fig. 2

模型预测结果

Bag-level:

$\text{Healthy}(I) = 1$

Instance-level:

$\text{Swelling}(x_1) = 1$

$\text{Recession}(x_2) = 0$

2. 与知识库对齐

将伪标签转换为知识库中的逻辑变量形式，使其可以代入领域知识规则进行推理。

对应的逻辑变量

$\text{Healthy}(I) = 1$

$\text{Swelling}(x_1) = 1$

$\text{Recession}(x_2) = 0$

…

3. 一致性检查

将逻辑变量代入知识库中的逻辑公式，统计满足的规则数量，判断一致性。

对应的逻辑规则（示例）

$\text{Healthy}(I) \rightarrow \neg\text{Swelling}(x)$

$\text{MMP}(I) \rightarrow \text{Swelling}(x)$

$\text{SP}(I) \rightarrow \text{Recession}(x)$

当前满足规则：6 / 10

4. 成本约束下修正

若与知识库不完全一致，在修改成本限制内，使满足的规则数量最大化。

示例修正

$\text{Swelling}(x_1) = 1$

与规则冲突

$\text{Swelling}(x_1) = 0$

修改成本：1（改 1 个变量）

5. 得到溯因标签

Abduced Labels

经修正后得到与知识库更一致的逻辑变量取值，用作模型的训练监督信号。

$\text{Healthy}(I) = 1$

$\text{Swelling}(x_1) = 0$

$\text{Recession}(x_2) = 0$

满足的规则数：9 / 10

优化目标（论文公式 4）

$$\max_{\Delta(\mathcal{O})} \text{Con}[\psi(f, \Delta(\mathcal{O}))] - \text{Con}[f]$$

$$\text{s.t. Cost}[\Delta(\mathcal{O})] \leq \mathcal{T}$$

一致性 Con(·)：满足的知识库公式数量，衡量预测与知识库的一致性。
修改成本 Cost(·)：调整变量的代价，通常与修改标签数量相关。
目标：在成本限制内修正标签，使满足的规则数量最大化。

核心思想

保留模型原始预测，利用专家知识进行小范围的合理修正。

注：6/10、9/10 和修改成本 1 为流程演示数值；配图：论文 Fig. 2（第 3 页）。

Abduction Process: Figure 4

论文 Figure 4 溯因推理过程

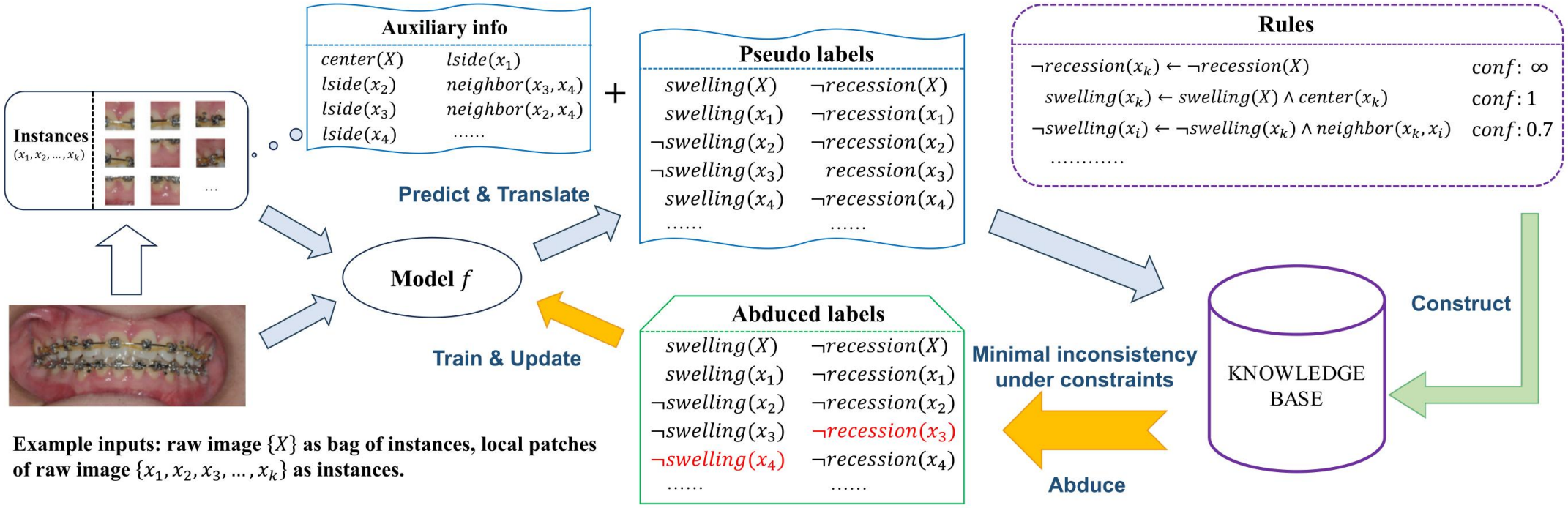


Fig. 4. The detailed illustration of the abduction process. The pseudo labels are first translated into logical variables and facts. Then we try to maximize the consistency between pseudo labels and the knowledge base by modifying some of the labels under cost constraints. For example, according to the logical rules, the bag-level annotation "no recession" means that all instances shall not show signs of recession, so the "recession" label of x_3 is modified. Also, according to the rules, since the instances around the center of the image are more likely to show signs of affection, as well as all neighbors show no signs of swelling, the label "swelling" of x_4 is modified, which is consistent with the ground truth.

Joint Optimization Objective（联合优化目标）

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

感知分类损失与知识一致性约束共同驱动 AB-MIML 训练

$$\mathcal{L} = -\sum_{n,c} \log P(Y_{n,c}|X_n) - \alpha \sum_{n,c,i} \log P(y^i_{n,c}|x^i_n) - \beta \sum_i \mathrm{Con}\left(\Delta(Y_n), KB_\theta\right)$$

Bag-level CE Loss

整张图像分类；监督来自医生标注

Instance-level CE Loss

局部 patch 分类；监督来自溯因标签

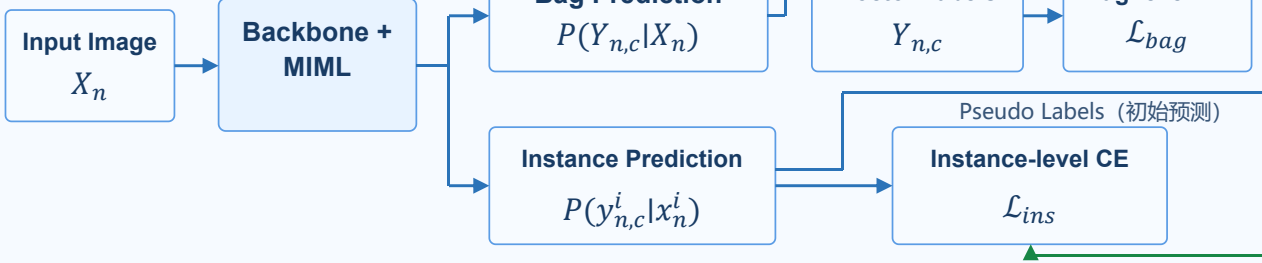
Knowledge Consistency

减少预测标签与知识库之间的不一致

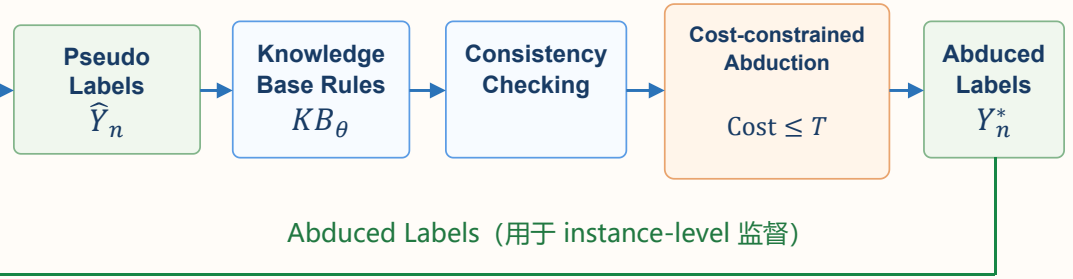
$\alpha = \alpha(t)$ 平衡两级分类损失

β 控制知识一致性约束强度

感知侧（Perception）



逻辑侧（Reasoning）



标签来源

Bag-level ground truth：来自医生标注

Instance-level 监督标签：由溯因修正得到

训练过程

模型侧：通过梯度下降最小化分类损失

逻辑侧：在成本约束下减少知识不一致

求解特点

溯因过程可视为 MAX-SAT 的变体

小规模问题：可用遍历方法

大规模问题：可采用启发式方法

数据集介绍

01 数据来源与基本信息

- **数据来源：** 南京市口腔医院
- **采集对象：** 正畸科患者口内图像
- **时间跨度：** 2017—2020 年
- **图像获取：** 数码相机拍摄（不同设备、不同分辨率）

02 类别设置

PH	GI	MMP	SP
Periodontal Health	Gingivitis	Mild to Moderate Periodontitis	Severe Periodontitis
牙周健康	牙龈炎	轻中度牙周炎	重度牙周炎

由牙周科专家完成四分类标注

03 数据集特点

- 仅提供图像级标签（image-level / bag-level）
- 不包含局部病灶 patch 的人工标注
- 实例级标签由伪标签和先验知识经溯因推理生成
- 随机划分训练 / 测试；按年份评估时间鲁棒性

04 数据集统计（Table 1）

训练集覆盖整体时间跨度；测试集按年份组织

类别	Train	Test			
		2017	2018	2019	2020
PH	1531	122	167	159	160
GI	1573	202	176	215	227
MMP	1575	189	183	189	183
SP	295	28	30	27	23

05 预处理与输入形式

原始口腔图像 → 牙龈轮廓识别与裁剪 → 裁剪聚焦牙周区域

Non-MIML 设置

统一缩放为 $3 \times 384 \times 384$

MIML 设置

每张图像分解为 10 个 patches

每个 patch 缩放为 $3 \times 128 \times 128$

裁剪后图像分辨率范围： $3 \times 800 \times 1200 \sim 3 \times 4000 \times 6000$

Experiments with Diverse Backbones

Table 2 实验结果

Table 2

Experimental results of different methods and models evaluated on several main metrics from 2017 to 2020. We assume that the labels of the original images are ground truth. The results are organized by groups of backbones and times.

Methods & Models	Metrics & Results											
	2017						2018					
	ACC	mPRE	mSEN	mSPE	mF1	mAUC	ACC	mPRE	mSEN	mSPE	mF1	MF1
VGG16	0.6987	0.7552	0.7297	0.8903	0.7374	0.8570	0.7104	0.7834	0.7428	0.8946	0.7584	0.8747
VGG16-MIML	0.7246	0.7753	0.7592	0.9000	0.7625	0.8824	0.7320	0.8026	0.7601	0.9026	0.7759	0.8625
VGG16-ABMIML	0.7320*	0.7805*	0.7797*	0.9026*	0.7763*	0.8874*	0.7446*	0.8118*	0.7697*	0.9071*	0.7859*	0.8920*
ResNet18	0.7782	0.8160	0.8110	0.9188	0.8116	0.9284	0.7770	0.8354	0.7961	0.9188	0.8109	0.9259
ResNet18-MIML	0.7911	0.8272	0.8254	0.9239	0.8233	0.9407	0.7896	0.8451	0.8126	0.9234	0.8247	0.9396
ResNet18-ABMIML	0.8133*	0.8426*	0.8449*	0.9325*	0.8406*	0.9446*	0.8094*	0.8595*	0.8349*	0.9306*	0.8437*	0.9462*
ResNet50	0.7930	0.8282	0.8189	0.9243	0.8209	0.9422	0.7968	0.8515	0.8252	0.9260	0.8341	0.9451
ResNet50-MIML	0.8152	0.8449	0.8460	0.9329	0.8425	0.9477	0.8129	0.8605	0.8380	0.9319	0.8463	0.9462
ResNet50-ABMIML	0.8318*	0.8577*	0.8573*	0.9384*	0.8558*	0.9510*	0.8291*	0.8713*	0.8506*	0.9378*	0.8587*	0.9524*
ResNeXt	0.7985	0.8328	0.8243	0.9264	0.8255	0.9407	0.7932	0.8462	0.8222	0.9248	0.8310	0.9410
ResNeXt-MIML	0.8189	0.8495	0.8380	0.9331	0.8418	0.9522	0.8201	0.8678	0.8439	0.9345	0.8521	0.9498
ResNeXt-ABMIML	0.8336*	0.8599*	0.8714*	0.9389*	0.8642*	0.9584*	0.8381*	0.8776*	0.8575*	0.9410*	0.8655*	0.9610*
ViT	0.8004	0.8349	0.8452	0.9270	0.8375	0.9436	0.7968	0.8481	0.8249	0.9261	0.8337	0.9447
ViT-MIML	0.8207	0.8501	0.8607	0.9344	0.8534	0.9541	0.8183	0.8618	0.8559	0.9340	0.8572	0.9510
ViT-ABMIML	**0.8429***	**0.8741***	**0.8777***	**0.9421***	**0.8748***	**0.9657***	**0.8399***	**0.8789***	**0.8728***	**0.9417***	**0.8740***	**0.9625***

Methods & Models	Metrics & Results											
	2019						2020					
	ACC	mPRE	mSEN	mSPE	mF1	mAUC	ACC	mPRE	mSEN	mSPE	mF1	mAUC
VGG16	0.7360	0.7883	0.7752	0.9028	0.7793	0.8886	0.7239	0.7765	0.7443	0.8973	0.7570	0.8870
VGG16-MIML	0.7411	0.7925	0.7786	0.9046	0.7833	0.8961	0.7391	0.7880	0.7557	0.9030	0.7685	0.8954
VGG16-ABMIML	0.7530*	0.8009*	0.7872*	0.9089*	0.7923*	0.9014*	0.7542*	0.8007*	0.7768*	0.9086*	0.7863*	0.9012*
ResNet18	0.7712	0.8153	0.8031	0.9158	0.8066	0.9356	0.7690	0.8122	0.7890	0.9142	0.7979	0.9342
ResNet18-MIML	0.7949	0.8338	0.8210	0.9244	0.8252	0.9462	0.7909	0.8296	0.8041	0.9222	0.8146	0.9464
ResNet18-ABMIML	0.8068*	0.8426*	0.8458*	0.9289*	0.8426*	0.9496*	0.8044*	0.8394*	0.8249*	0.9273*	0.8305*	0.9514*
ResNet50	0.7932	0.8317	0.8270	0.9247	0.8267	0.9401	0.7926	0.8312	0.8284	0.9233	0.8270	0.9412
ResNet50-MIML	0.8102	0.8451	0.8573	0.9312	0.8462	0.9479	0.8094	0.8438	0.8389	0.9293	0.8398	0.9486
ResNet50-ABMIML	0.8288*	0.8598*	0.8701*	0.9378*	0.8625*	0.9537*	0.8280*	0.8568*	0.8547*	0.9366*	0.8538*	0.9555*
ResNeXt	0.7966	0.8344	0.8293	0.9259	0.8294	0.9398	0.7943	0.8315	0.8316	0.9247	0.8275	0.9377
ResNeXt-MIML	0.8186	0.8514	0.8483	0.9340	0.8470	0.9486	0.8044	0.8392	0.8382	0.9282	0.8354	0.9429
ResNeXt-ABMIML	0.8356*	0.8651*	0.8766*	0.9401*	0.8684*	0.9597*	0.8381*	0.8639*	0.8629*	0.9407*	0.8611*	0.9588*
ViT	0.7983	0.8346	0.8406	0.9263	0.8352	0.9418	0.8061	0.8400	0.8409	0.9292	0.8362	0.9441
ViT-MIML	0.8186	0.8510	0.8650	0.9340	0.8549	0.9477	0.8111	0.8439	0.8539	0.9311	0.8451	0.9462
ViT-ABMIML	**0.8458***	**0.8727***	**0.8836***	**0.9436***	**0.8764***	**0.9640***	**0.8465***	**0.8707***	**0.8781***	**0.9438***	**0.8724***	**0.9690***

表示组内最优，粗体表示全局最优。表格图片直接截自论文第 8 页 Table 2。

01 实验设置

- 对比骨干：VGG16、ResNet18、ResNet50、ResNeXt、ViT
- 设置：Baseline、MIML、ABMIML 三种方案
- 每个 epoch 留出约 1/10 训练数据用于验证
- 超过 5 个 epoch 无显著提升时提前停止训练

02 评价指标与对比

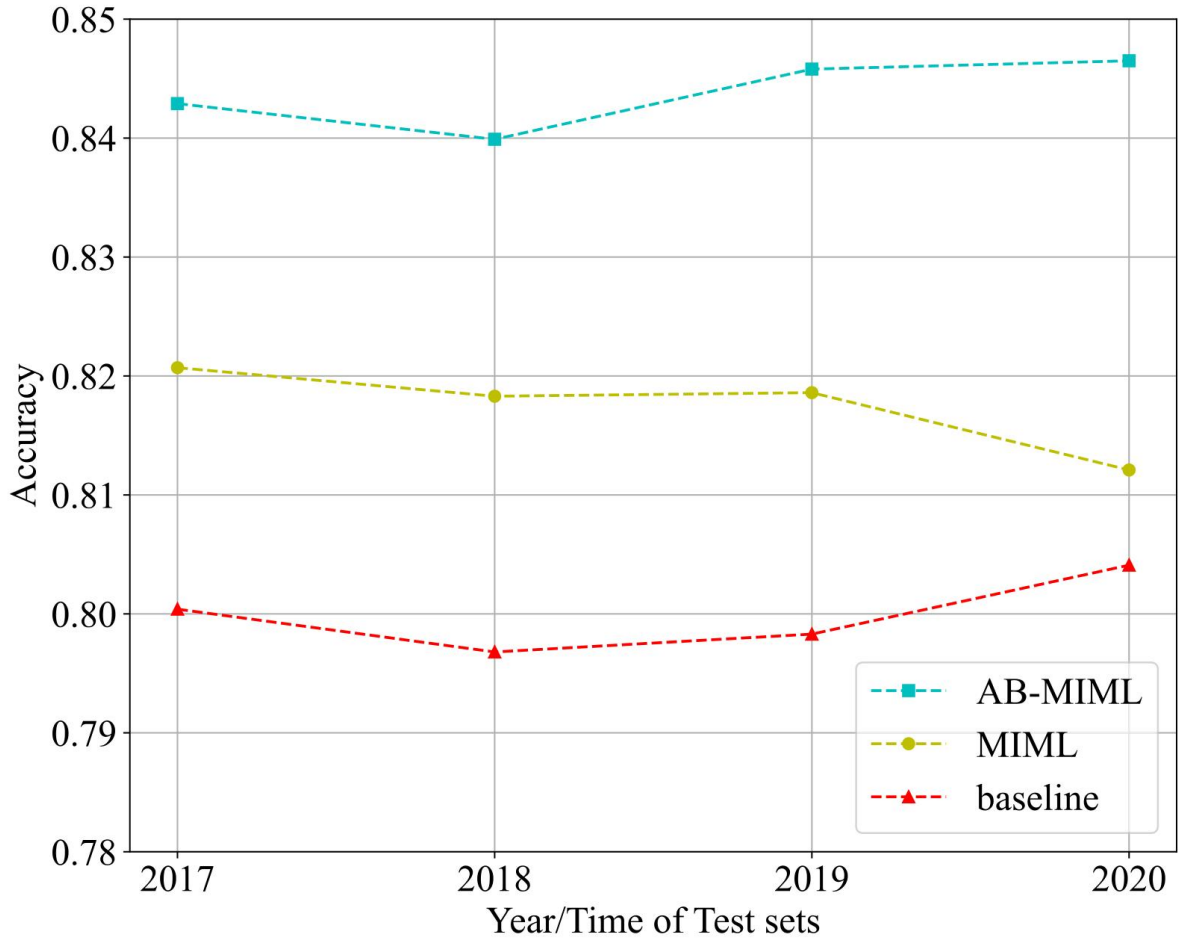
- 指标：ACC、mPRE、mSEN、mSPE、mF1、mAUC
- 在同一 backbone、同一测试年份内纵向比较
- Baseline → MIML：考察建模增益
- MIML → ABMIML：考察溯因推理增益

03 结论一：整体性能优势

- 几乎所有评价指标均优于 baseline
- MIML 建模优于原始图像直接分类
- 引入溯因推理后，各骨干性能进一步提升
- 纵向消融验证建模与先验知识的贡献

Experiments with Diverse Backbones

01 跨年份稳定性分析



02 实验设置

- 按年份组织测试集：2017—2020
- 以 ViT 骨干网络组为例进行分析
- 对比 baseline、MIML、AB-MIML

03 关键观察

- 跨年份波动较小，保持稳定表现。
- 整体准确率约为 84%—85%。
- 相比 baseline 提升约 5 个百分点。

04 结论二：时间鲁棒性

在 2017—2020 年的测试集上，
AB-MIML 持续优于对照方法，
体现出较好的跨年份稳定性。

Experiments with Detection Models

Table 3 与目标检测方法的对比

Table 3
A brief exhibition of experimental results of our proposed method compared with the detection models. The backbone models in the experiment are all implemented using the well-established versions.

Methods & Models	Metrics & Results											
	SampleNum = 1200						SampleNum = 2000					
	ACC	mPRE	SEN_{PH}	SEN_{GI}	SEN_{MMP}	SEN_{SP}	ACC	mPRE	SEN_{PH}	SEN_{GI}	SEN_{MMP}	SEN_{SP}
RCNN	0.7121	0.7418	0.9615	0.4151	0.6600	0.8372	0.7240	0.7510	0.9600	0.3846	0.7200	0.8750
F-RCNN	0.6959	0.7220	0.9423	0.3725	0.6531	0.8333	0.7273	0.7510	0.9231	0.3962	0.7551	0.8636
YOLOv1	0.7231	0.7463	0.8846	0.4118	0.7551	0.8605	0.7323	0.7568	0.9038	0.4340	0.7708	0.8444
YOLOv7	0.7474	0.7730	0.9423	0.4400	0.7755	0.8372	0.7577	0.7797	0.8846	0.5000	0.7959	0.8605
ResNet50-ABMIML	0.8041	0.8183	0.7692	0.7800	0.7959	0.8837	0.8196	0.8318	0.8077	0.8000	0.7959	0.8837
ViT-ABMIML	**0.8247***	0.8339	0.8077	0.8000	0.7755	0.9302	**0.8367***	0.8445	0.8077	0.8400	0.7843	0.9302

01 公平对比设置

数据：1,200 / 2,000 张精细标注图像，
类别大致均衡；所有方法重新评估。

对照：RCNN、Faster-RCNN、
YOLOv1、YOLOv7 的成熟实现。

测试：仅保留置信度 > 0.5 的框，
依据候选框类别确定整图标签。

02 总体准确率优势

ViT-ABMIML 在两组数据上 ACC 最高

1,200 张	2,000 张
82.47%	**83.67%**
+7.73 个百分点	**+7.90 个百分点**

相对最佳检测基线 YOLOv7 的 ACC 增益

03 类别表现与原因

主要差异：检测方法对 GI 的召回偏低

GI 召回率（1,200 / 2,000 张）
ViT-ABMIML：80% / 84%
检测方法：约 37%—44% / 38%—50%

作者推测：牙龈肿胀的纹理差异细微，
可能使检测方法将 GI 误判为 PH。

The Impact of Pretraining

Table 4 鲁棒性分析：预训练的影响

Table 4
Experimental results of pretrained and non-pretrained models. To be brief, we select two representative backbones to demonstrate the effectiveness of our approach. The 'NaN' in the table indicates that the combination of methods can hardly converge to a feasible solution.

Methods & Models	Metrics & Results									
	Pretrained					Non-pretrained				
	ACC	mPRE	mSEN	mSPE	mF1	ACC	mPRE	mSEN	mSPE	mF1
ResNet50	0.7906	0.8265	0.8194	0.9277	0.8214	0.7345	0.7554	0.7579	0.9024	0.7562
ResNet50-MIML	0.8164	0.8389	0.8377	0.9288	0.8362	NaN	NaN	NaN	NaN	NaN
ResNet50-ABMIML	0.8308*	0.8584	0.8562	0.9342	0.8563	0.8178*	0.8372	0.8364	0.9162	0.8388
ViT	0.8006	0.8368	0.8387	0.9273	0.8374	0.7534	0.7722	0.7753	0.9062	0.7748
ViT-MIML	0.8165	0.8514	0.8615	0.9327	0.8560	NaN	NaN	NaN	NaN	NaN
Vit-ABMIML	**0.8432***	0.8683	0.8796	0.9415	0.8728	**0.8207***	0.8427	0.8462	0.9356	0.8452

01 实验设置

固定骨干：ResNet50、ViT；
分别比较预训练与非预训练设置。

测试集随机抽取，不再按年份组织；
SP 类约抽取 20%，其余类约 10%。

NaN：200 epochs 后未收敛，
或退化为对所有输入预测同一标签。

02 非预训练：改善收敛

两种骨干的 MIML 均出现 NaN；
加入 AB-MIML 后得到有效分类结果。

ResNet50	ViT
81.78%	**82.07%**
+8.33 个百分点	**+6.73 个百分点**

ACC 增益相对于对应的非预训练基础模型

03 预训练：进一步提升

AB-MIML 的 ACC 在组内最高

ResNet50：83.08%（较 MIML +1.44 pp）
ViT：84.32%（较 MIML +2.67 pp）

先验知识辅助收敛，也能提升预训练模型表现。

The Impact of Noisy Datasets

Table 5 鲁棒性分析：标签噪声的影响

Table 5
A brief overview of experimental results of different methods evaluated on several key metrics of test datasets. We assume that there is a certain ratio of noise in the bag-level annotation. The methods evaluated using a pretrained backbone are identified by the suffix "(P)". The "NaN" in the table indicates that the model can hardly converge to a feasible solution under such settings.

Methods & Models	Metrics & Results											
	$\eta = 0\%$			$\eta = 5\%$			$\eta = 10\%$			$\eta = 20\%$		
	ACC	mSEN	mF1	ACC	mSEN	mF1	ACC	mSEN	mF1	ACC	mSEN	mF1
ResNet50	0.7345	0.7579	0.7562	0.7300	0.7602	0.7696	0.7024	0.7464	0.7523	0.6605	0.7164	0.7153
ResNet50-MIML	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ResNet50(P)-MIML	0.8164	0.8377	0.8362	0.8090	0.8254	0.8266	0.7588	0.7893	0.7830	0.7207	0.7536	0.7601
ResNet50-ABMIML	0.8178	0.8364	0.8388	0.8120	0.8298	0.8314	0.7541	0.7842	0.7793	0.7157	0.7636	0.7614
ResNet50(P)-ABMIML	**0.8308***	0.8562	0.8563	**0.8280***	0.8423	0.8507	**0.7852***	0.8040	0.8121	**0.7460***	0.7670	0.7614

01 标签噪声设置

骨干：ResNet50；(P) 表示采用预训练。
将医生原始标签视为无噪声基准。

噪声率：0%、5%、10%、20%；
随机抽取训练样本，将标签改为其他类。

SP 样本较少，需控制向该类注入噪声，
避免其实际噪声比例过高。

02 20% 噪声下的结果

同为预训练设置，比较 ACC

AB-MIML (P)	MIML (P)
74.60%	72.07%

AB-MIML 高出 2.53 个百分点

03 观察与解释

预训练 AB-MIML 在各噪声率下 ACC 最高；
噪声增大时，性能仍会下降。

非预训练 MIML 均为 NaN，
提示该设置下难以有效收敛。

作者解释：专家先验提供约束，
有助于抑制错误标签、实现一定自纠正。

Robustness across Classes

Table 6 鲁棒性分析：不同类别的表现

Table 6
Experimental results for class-wise analysis. The models used in the experiment are all pretrained.

Methods & Models	Metrics & Results								
	ACC	PRE_{PH}	PRE_{GI}	PRE_{MMP}	PRE_{SP}	SEN_{PH}	SEN_{GI}	SEN_{MMP}	SEN_{SP}
ResNet50	0.7877	0.7290	0.7742	0.8559	0.8529	0.8721	0.6829	0.8112	0.8286
ResNet50-MIML	0.8091	0.7480	0.7813	0.8819	0.9412	0.8676	0.7114	0.8394	0.9143
ResNet50-ABMIML	0.8344*	0.8000	0.7851	0.9043	0.9459	0.8767	0.7724	0.8353	1.0000
ViT	0.7984	0.7364	0.7727	0.8697	0.9394	0.8676	0.6911	0.8313	0.8857
ViT-MIML	0.8171	0.7610	0.7946	0.8782	0.9444	0.8721	0.7236	0.8394	0.9714
Vit-ABMIML	**0.8438***	0.7935	0.8069	0.9106	1.0000	0.8950	0.7642	0.8594	0.9714

指标说明：ACC 为总体准确率；PRE 为分类别精确率；SEN 为分类别召回率。

01 类别不平衡与设置

SP 训练样本仅 295 张；
其余三类约 1,500 张（见 Table 1）。

采用 ResNet50、ViT 两种骨干，
本实验所有模型均经过预训练。

比较 Baseline、MIML、AB-MIML，
逐类分析 PH、GI、MMP、SP。

02 SP 类：少样本仍较易识别

AB-MIML 在 SP 类上的表现突出

ResNet50 • SP 召回率
100%

ViT • SP 精确率
100%

作者解释：SP 的牙根暴露等局部特征
较明显，可能降低了类别区分难度。

03 整体提升与边界

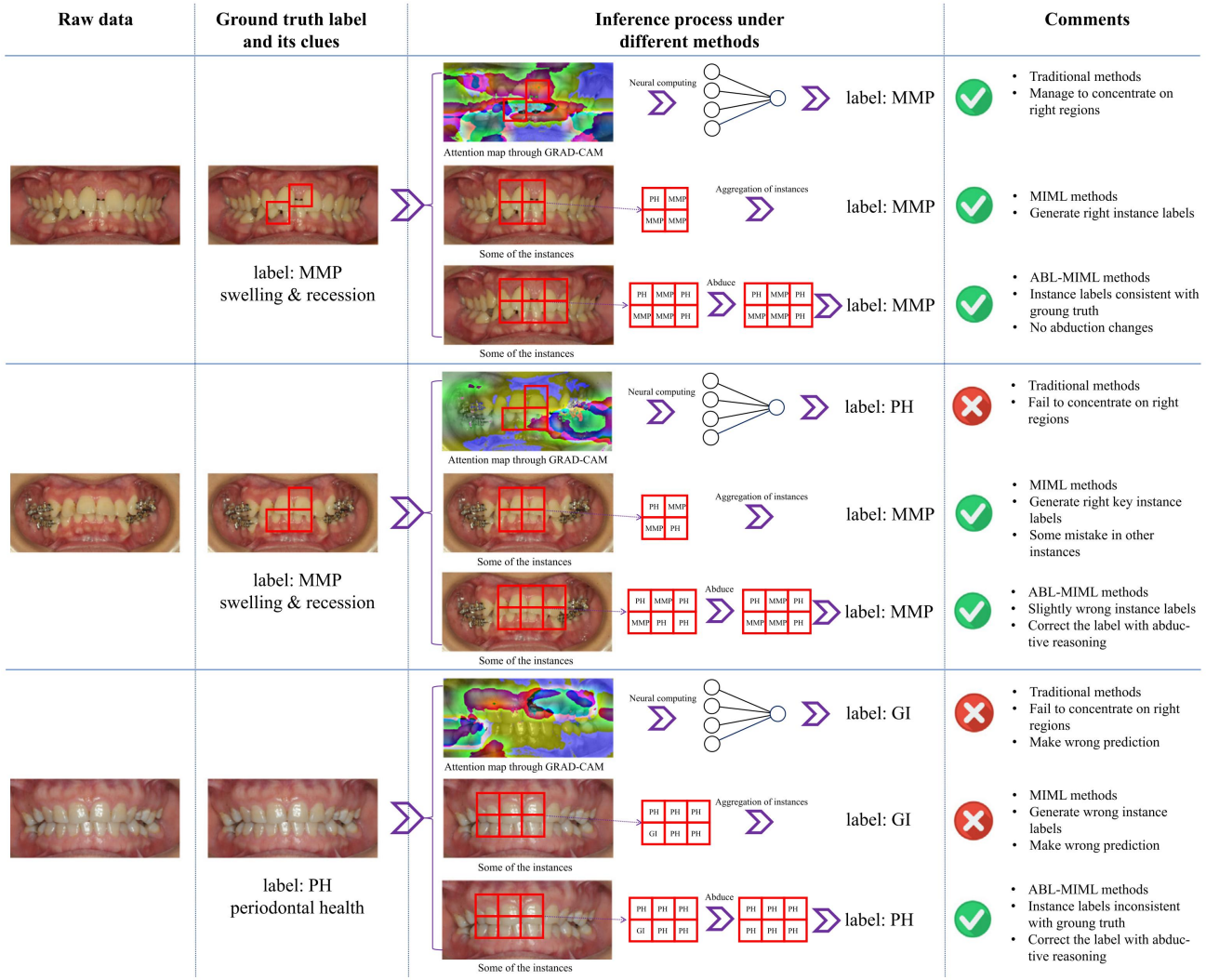
总体 ACC：83.44% / 84.38%
分别对应 ResNet50 / ViT，均为组内最高。

相对基础模型，多数分类别指标提升；
SP 优势突出，PH 精确率也有所改善。

并非逐项最优：ResNet50 的 MMP 召回率
略低于 MIML（83.53% vs. 83.94%）。

Effectiveness of Abductive Methods

Figure 7 典型案例与 Grad-CAM 可视化



01 MMP：三种方法均预测正确

Baseline、MIML、AB-MIML 均输出 MMP。
局部标签与知识一致，无需溯因修改。

02 MMP：局部建模纠正整图误判

Baseline 关注错误区域，将 MMP 误判为 PH；
MIML 聚合关键实例后正确输出 MMP；
AB-MIML 进一步修正部分局部错误标签。

03 PH：溯因推理纠正局部错误

Baseline 和 MIML 均误判为 GI。
AB-MIML 发现局部肿胀预测与知识库冲突，
在成本约束下修正标签，最终输出 PH。

→ 从局部感知到知识约束纠错

局部预测 → 一致性检查 → 成本约束修正

实例置信度参与标签翻转成本计算；
修正结果应更符合专家知识约束。