主办 CCF计算机视觉专业委员会

COMPUTER VISION NEWSLETTER

CCF 计算机视觉专委会简报

02 2024

总第 40 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2024 年第 02 期

总第 40 期

主　办　CCF 计算机视觉专业委员会

CCF 计算机视觉专委会

编委会

CONTENTS
简报目录

CCF 计算机视觉专委会
CCFCV.CCF.ORG.CN
CCFCVN@GMail.com

CCF-CV 走进高校系列报告会

第 136 期 兰州理工大学

2024 年 4 月 17 日下午，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、兰州理工大学承办的第 136 期 CCF-CV 走进高校系列报告会——“智能视觉感知与计算”在兰州理工大学举行。本期报告会邀请南京大学**吴建鑫**教授、合肥工业大学**贾伟**教授、西北工业大学**夏勇**教授、中国科学院计算技术研究所**王瑞平**研究员四位专家学者做特邀报告。兰州理工大学计算机与通信学院院长**冯涛**、副院长**李晓旭**担任本次报告会的执行主席。

活动首先由冯涛教授致欢迎辞，首先对所有参与此次CCF-CV走进高校系列报告会的专家学者表示最诚挚的欢迎和衷心的感谢。冯教授强调，这是推动学院科研水平、拓宽师生学术视野的重要机会，同时期望此次报告会能够加深与各位专家学者的交流与合作，丰富学院的学术氛围，推动学科的发展。随后吴建鑫教授、夏勇教授、贾伟教授、王瑞平研究员分别做主题报告。最后，李晓旭教授进行了活动总结，感谢四位专家的精彩报告和师生们的热情参与，并对 CCF-CV 专委会和学校为本次活动顺利开展提供的支持表示感谢。祝贺本次活动取得了圆满成功！

第 137 期 苏州科技大学

2024 年 5 月 20 日，由中国计算机学会计算机视觉专委会 (CCF-CV) 主办、苏州科技大学承办的第 137 期走进高校系列报告会以线上线下相结合的方式在苏州科技大学成功举办。出席本次活动的嘉宾有山西大学**梁吉业**教授、浙江大学**李玺**教授、中山大学**郑伟诗**教授、中国科学院自动化研究所**姚涵涛**副研究员、哈尔滨工业大学**洪晓鹏**教授、南京大学**霍静**副教授、南京理工大学**杨杨**教授、北京理工大学**李爽**副教授和西北工业大学**梁国强**副教授。苏州科技大学校长**顾菊平**教授与苏州科技大学副校长**肖洋**教授参加了本次报告会。本次会议执行主席为苏州科技大学电子与信息工程学院院长**胡伏原**教授，CCF 苏州分部主席**龚声蓉**教授、苏州科技大学讲师**程涵婧**、江南大学**赵少川**博士。

活动首先由肖洋教授为报告会致欢迎词，肖洋教授对各位专家学者的到来表示热烈的欢迎和衷心的感谢，并预祝本次活动取得圆满成功。随后梁吉业教授、李玺教授、郑伟诗教授、姚涵涛副研究员、洪晓鹏教授、霍静副教授、杨杨教授、李爽副教授、梁国强副教授分别做主题报告。本次会议除线下报告外，同时通过 CCF 计算机视觉专委会官方账号在 B 站进行腾讯会议的直播，

总观看人数超过 200 余人。参加本次活动的老师和同学认真聆听了报告，并与报告嘉宾热情地交流与互动，共同探讨学术内容。在活动的尾声，执行主席胡伏原教授上台发表了总结致辞，感谢各位专家的精彩报告以及线上线下师生们的热情参与，最后祝贺本次活动取得了圆满成功！

责任编委 李实英 黄岩

CCF-CV 走进企业系列交流会

第 28 期 超集信息

2024 年 5 月 31 日，CCF 计算机视觉专委会联合超集信息会共同举办"生成式大模型研究趋势及其高效训练技术交流会"，南京大学**王利民**教授、华中科技大学**王兴刚**教授、浙江大学**赵洲**教授、哈尔滨工业大学**左旺孟**教授等出席本次活动带来最新研究成果、技术突破及未来趋势分享，旨在共同助推人工智技术创新与进步。

会议开场，苏州超集信息科技有限公司销售及市场副总裁廖治国指出随着生成式 AI 大模型在多个领域的创新应用，我们日常生活迎来了前所未有的变革。但随着科研深入，算力资源供给和日常使用问题日益凸显，成为人工智能实现普惠的最大难点。面对科研等多场景用户的迫切算力需求，超集信息将持续输出更高效、更稳定、更绿色的高性能计算解决方案，助力更多应用落地，共同助推科技进步。随后，王利民教授、王兴刚教授、赵洲教授、左旺孟教授分别围绕多模态研究和生成模型进行了研究报告。最后，超集信息资深售前工程师沈佳威带来了全方位算力解决方案分享，助力高效应对日益复杂的计算挑战。

本场交流会上，通过前沿技术和理念的交流，与会嘉宾明确了未来人工智能技术发展的方向。随着算力资源的不断优化和多模态技术的深入应用，人工智能正朝着更加智能化、个性化的方向发展，为我们的生活带来无限可能。

未来，超集信息也将和 CCF 计算机视觉专委会继续保持深度合作，以丰富的高性能计算解决方案经验及专业技术，助力多模态等人工智能技术高质量发展，携手共创人工智能美好未来。

责任编委 潘金山 朱安娜

CCF-CV 秘书处 2024 年度第一次工作会议顺利召开

2024 年 4 月 7 日，中国计算机学会计算机视觉专委会 (CCF-CV) 秘书处年度第一次工作会议于北京召开。专委会副主任王亮研究员参会指导工作，秘书处全体成员参加了会议，会议由秘书长王瑞平研究员主持。本次会议主要围绕秘书处未来工作规划和后续活动改进方案等议题展开讨论。

会议首先回顾并总结了秘书处过去几年的工作情况，并针对当前的工作需求，对成员的职责进行了重新分配和优化。为了确保常规活动和其他特殊活动的高效执行，会议提出了实施轮岗制度和 AB 角负责机制，以此来保障活动的连续性和稳定性。随后，会议围绕“创精品、务实效、惠全域”的主题，深入讨论了如何提高各类品牌活动的质量、提升委员们参与/协助组织专委活动的积极性等议题。具体包括增加走进高校活动的多样性；以企业需求为导向、有针对性地组织走进企业的交流活动；提供不同形式的宣传素材并扩大专委的宣传力度；年度学术研讨会 RACV 的组织等。会议最终形成了一套具体可行的执行方案，旨在通过这些措施和改革，进一步提升秘书处的工作质量和效率。

新一届秘书处成员将保持始终如一的工作热情，致力于确保各项活动的高效组织与成功执行，增强专委会的凝聚力，推动专委会的整体发展。

责任编委 朱安娜

中国计算机学会计算机视觉专委会（CCF-CV）

2024 年执行委员增选申请开始啦！

自 2013 年 10 月成立以来，中国计算机学会（CCF）计算机视觉专业委员会（ccfcv.ccf.org.cn）发展迅速，举办了很多有影响力的活动，如计算机视觉前沿进展研讨会（RACV）、CCF-CV 走进高校系列报告会、CCF-CV 走进企业系列交流会、CCF-CV 视界无限系列研讨会、计算机视觉前沿讲习班，与中国自动化学会模式识别与机器智能专委会、中国图象图形学学会视觉大数据专委会、中国人工智能学会模式识别专委会共同举办中国模式识别与计算机视觉大会（PRCV），定期出版专委简报，建设专委中英文网站，专委微信公众号文章平均阅读上千次，专委活动视频在专委 Bilibili 账号（https://live.bilibili.com/22339632）发布。搭建了全方位、高水平、大规模的计算机视觉领域交流平台。专委会成立 10 年以来，在 CCF 专委评估中获得“特色活动奖”、“综合进步奖”、“优秀专委奖”、“年度特别奖”等 7 个奖项。为了保持专委会的活力、促进国内外视觉领域人员的交流和合作，专委会现开放 2024 年计算机视觉专委会的执行委员增选工作。

一、申请时间

2024 年 6 月 10 日— 2024 年 9 月 10 日。

二、申请流程

填写申请表（点击最下方阅读原文可直接下载），发送给秘书处（ccfcv@139.com），主题“2024 新执行委员申请-姓名-单位”。（注：推荐人必须是现任专委执行委员，名单可以从专委网站查询。电子版申请表中需填写推荐人姓名和意见，执行委员增选成功后可以补签签名）。

三、申请资格

任职国内外学术界或企业界副教授或等同级别以上的人员，拥有计算机视觉相关领域的高水平研究成果，是 CCF 计算机视觉专委委员，且积极参加计算机学会计算机视觉专委会的各项活动。特别优秀的讲师、企业人士亦可考虑。

四、申请需求

现任专委执行委员每人可推荐最多 3 名候选人。本次申请结果将在 “2024 年中国模式识别与计算机视觉大会” （http://www.prcv.cn）期间（2024 年 10 月 18 日-20 日）举行的专委工作年会上投票确定（申请者届时必须“注册参会”）。

五、特别说明

按照 CCF 的新规定，CCF 专业会员通过 CCF 会员系统关注相关专委后即加入专委并成为其委员，其后每年可以更改一次关注的专委。委员在专委中无选举权和被选举权，但具有对专委的评价权。原来的专委委员自动升级为专委执行委员，享有选举权、被选举权以及对专委的评价权。

责任编委 毋立芳 任传贤

专题综述

视觉 Mamba：面向高效视觉表示学习的双向状态空间模型

朱良辉[1] 廖本成[1] 张骞[2] 王鑫龙[3] 刘文予[1] 王兴刚[1]
[1]华中科技大学 [2]北京地平线信息技术有限公司 [3]北京智源人工智能研究院

本文是华中科技大学、地平线公司与北京智源人工智能研究院团队合作研究的成果，发表在ICML 2024的工作Vision Mamba[1]。论文研究的问题是如何设计新型神经网络来实现高效的视觉表示学习。该任务要求神经网络模型能够在处理高分辨率图像时既保持高性能，又具备计算和内存的高效性。先前的方法主要依赖自注意力机制来进行视觉表示学习，但这种方法在处理长序列时速度和内存使用上存在挑战。论文提出了一种新的通用视觉主干模型Vision Mamba，简称Vim，该模型使用双向状态空间模型（SSM）对图像序列进行位置嵌入，并利用双向SSM压缩视觉表示。在ImageNet[2]分类、COCO[2]目标检测和ADE20K[3]语义分割任务中，Vim相比现有的视觉Transformer[4]（如DeiT[5]）在性能上有显著提升，同时在计算和内存效率上也有显著改进。例如，在进行分辨率为1248×1248的批量推理时，Vim比DeiT快2.8倍，GPU内存节省86.8%。这些结果表明，Vim能够克服在高分辨率图像理解中执行Transformer样式的计算和内存限制，具有成为下一代视觉基础模型主干的潜力。此外，我们还将介绍 Vim 的后续工作——Vision GLA（ViG） 和 Diffusion GLA（DiG）。ViG和DiG利用带有 FlashAttention 优化的门控线性注意力来构建图像理解和图像生成模型，在精度和速度上相对于Vim模型取得了进一步的突破。

一、研究背景

图像表示学习是计算机视觉领域的重要研究课题，其目的是通过模型学习从图像中提取有意义的特征，从而应用于各种视觉任务中。目前，视觉 Transformer（Vision Transformer，ViT[4]）和卷积神经网络（Convolutional Neural Networks, CNNs）是图像表示学习中最常用的方法。然而，这些方法在理论上存在一些局限性。

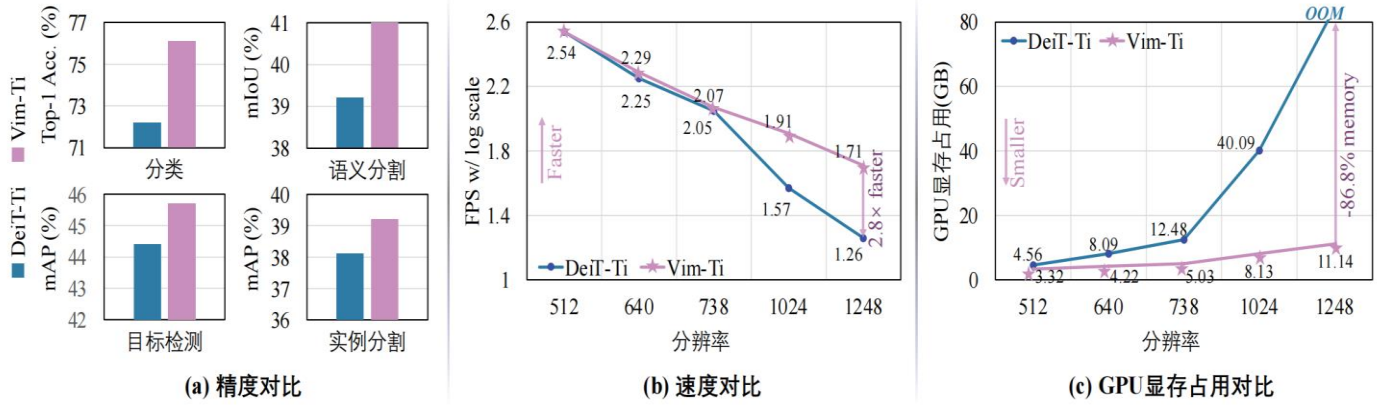


图 1　本文所提出的 Vision Mamba (Vim)和基于 Transformer 的 DeiT 模型进行精度与效率对比：Vim 在图像分类、目标检测、语义分割、实例分割任务上获得了更好的精度，且在高清分辨率图像处理上呈现出巨大的优势。

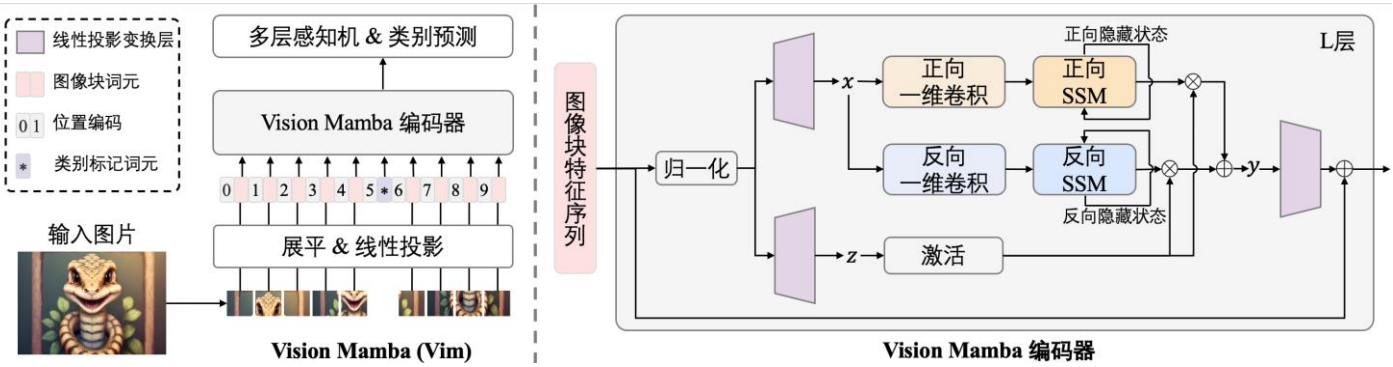


图 2　本文所提出的 Vim 模型的网络构架图

视觉 Transformer 利用自注意力机制能够取得全局的感受野，在大规模自监督预训练和下游任务中表现出色，但其自注意力机制在处理长序列依赖和高分辨率图像时，带来了计算和内存的巨大开销。具体而言，自注意力机制的计算复杂度是输入的图像块序列长度的平方，这使得其在处理高分辨率图像时非常耗时且占用大量内存。尽管一些研究提出了改进方法，如窗口注意力机制[6,7]，这些方法虽然降低了复杂度，但导致感受野被局限在局部的窗口内部，失去了原本全局感受野的优势。

另一方面，卷积神经网络在处理图像时，通过使用固定大小的卷积核来提取局部特征。然而，卷积神经网络在捕捉全局上下文信息方面存在局限性，因为卷积核的感受野是有限的，虽然一些研究引入了金字塔结构或大卷积核来增强全局信息提取能力，但这些改进仍然无法完全克服 CNN 在处理长序列依赖方面的不足。

在自然语言处理领域，Mamba[11]方法的出现给高效率长序列建模带来了很好的发展契机。Mamba 是状态空间模型（state space model, SSM）方法的最新演进。Mamba 提出了一种输入自适应的状态空间模型，能够更高质量地完成序列建模任务。与此同时，该方法在处理长序列建模问题时有着次二次方的复杂度与更高的处理效率。然而，Mamba 方法并不能够直接应用于视觉表征学习，因为 Mamba 方法是为自然语言领域的因果建模而设计的，它缺少对于二维空间位置的感知能力以及缺少全局的建模能力。

为了克服上述 Transformer 和 CNN 的理论局限性，启发于自然语言处理领域 Mamba 的成功，本文提出了一种新的通用视觉主干模型——Vision Mamba (Vim)。该模型基于状态空间模型[10]（State Space Models, SSMs），利用其在长序列建模中的高效性，提供了一种新的视觉表示学习方法。该模型提出了双向状态空间模型来适配视觉特征的多方向性，并引入位置编码来针对图像单元进行标记。本文提出的 Vim 模型通过双向 SSM 对图像序列进行位置嵌入和压缩，不仅在 ImageNet 分类任务上表现出色，还在 COCO 目标检测和 ADE20K 语义分割任务中展示了优异的性能。与现有的视觉 Transformer 如 DeiT 相比，Vim 在计算和内存效率上有显著提升。

此外，为了进一步推动高效视觉表征学习在图像识别和图像生成领域的应用，我们还开发了 Vision GLA（ViG[16]）和 Diffusion GLA（DiG[17]）获得了更加高效率的图像理解、图像生成表征学习建模。

ViG 通过引入门控线性注意力机制，极大地提升线性视觉序列学习的高效性，获得比 Vision Mamba 更好的性能，不仅在高清图像上获得全局感受野和线性计算量带来的效率优势，在中小图像上也展示出了和高度优化的 Transformer 和 CNN 模型相当的效率。

DiG 首次提出层级异向的因果建模方式，并以此构建了高效高质量的图像生成模型方法。该方法通过简洁高效的算法实现，在不同模型尺寸，不同图像大小下都展示出了相较于基线模型更强的生成效果与效率。在长序列下的效率甚至超过了一众次二次方复杂的的扩散模型。

二、Vision Mamba方法介绍

1．前言：状态空间模型

状态空间模型，比如结构化状态空间序列模型[10]

（S4）和 Mamba[11]是启发于连续系统，该系统通过隐藏状态$h(t) \in \mathbb{R}^N$将一维函数或序列$x(t) \in \mathbb{R}$映射到$y(t) \in \mathbb{R}$。该系统使用$A \in \mathbb{R}^{N\times N}$作为演化参数，并使用$B \in \mathbb{R}^{N\times 1}$和$C \in \mathbb{R}^{1\times N}$作为投影参数。连续系统的工作方式如下：

$$h'(t) = Ah(t) + Bx(t)$$

$$y(t) = Ch(t)$$

S4 和 Mamba 是连续系统的离散版本，它们包含一个时间尺度参数Δ，用于将连续参数A和B转换为离散参数$\bar{A}$和$\bar{B}$。常用的方式是零阶保持，其定义如下：

$$\bar{A} = \exp(\Delta A)$$

$$\bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B$$

将$\bar{A}$和$\bar{B}$离散化后，使用步长Δ的离散版本可以重写为：

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t$$

$$y_t = Ch_t$$

最后，模型可以使用全局的卷积来并行的计算：

$$\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \dots, C\bar{A}^{M-1}\bar{B})$$

$$y = x * \bar{K}$$

其中M是输入序列x的长度，$\bar{K} \in \mathbb{R}^M$是结构化的卷积核。

2．Vision Mamba 结构

所提出的 Vision Mamba 如图 1 所示。标准的 Mamba 模块是为一维的文本序列所设计的。为了适配视觉信号，我们首先将二维图像$t \in \mathbb{R}^{H\times W\times C}$转换为展平的二维图像块序列$x_p \in \mathbb{R}^{J\times(P^2\cdot C)}$，其中$(H, W)$是输入图像的尺寸，$C$是通道数，$P$是图像块的尺寸。接下来，我们将$x_p$线性投影到大小为$D$的向量，并添加位置编码$E_{pos} \in \mathbb{R}^{(J+1)\times D}$，如下所示：

$$T_0 = [t_{cls}; t_p^1 W; \dots; t_p^J W] + E_{pos}$$

其中t_p^j是t中的第j个图像块，$W \in \mathbb{R}^{(P^2\cdot C)\times D}$是可学习的投影变换矩阵。受 ViT[4]的启发，我们也使用类别标记t_{cls}来表示整个图像块序列。然后，我们将标记序列T_{l-1}输入到 Vim 编码器的第l层，并得到输出T_l。最后我们对输出类别标记T_L^0进行归一化，并将其送入多层感知机（MLP）分类头以获得最终类别预测$\hat{p}$：

$$T_l = Vim(T_{l-1}) + T_{l-1}$$

$$f = Norm(T_L^0)$$

$$\hat{p} = MLP(f)$$

其中 Vim 是提出的视觉 Mamba 模块，L是层数，$Norm$是归一化层。

算法 1：Vim 模块流程

输入：图像块序列T_{l-1}

输出：图像块序列T_l

/* normalize the input sequence $\mathbf{T}'_{l-1}$ */
$\mathbf{T}'_{l-1}$: (B, M, D) ← $\mathbf{Norm}(\mathbf{T}_{l-1})$
$\mathbf{x}$: (B, M, E) ← $\mathbf{Linear}^{\mathbf{x}}(\mathbf{T}'_{l-1})$
$\mathbf{z}$: (B, M, E) ← $\mathbf{Linear}^{\mathbf{z}}(\mathbf{T}'_{l-1})$
/* process with different direction */
for o in {forward, backward} **do**
  $\mathbf{x}'_o$: (B, M, E) ← $\mathbf{SiLU}(\mathbf{Conv1d}_o(\mathbf{x}))$
  $\mathbf{B}_o$: (B, M, N) ← $\mathbf{Linear}_o^{\mathbf{B}}(\mathbf{x}'_o)$
  $\mathbf{C}_o$: (B, M, N) ← $\mathbf{Linear}_o^{\mathbf{C}}(\mathbf{x}'_o)$
  /* softplus ensures positive $\mathbf{\Delta}_o$ */
  $\mathbf{\Delta}_o$: (B, M, E) ← $\log(1 + \exp(\mathbf{Linear}_o^{\mathbf{\Delta}}(\mathbf{x}'_o) + \mathbf{Parameter}_o^{\mathbf{\Delta}}))$
  /* shape of $\mathbf{Parameter}_o^{\mathbf{A}}$ is (E, N) */
  $\overline{\mathbf{A}_o}$: (B, M, E, N) ← $\mathbf{\Delta}_o \otimes \mathbf{Parameter}_o^{\mathbf{A}}$
  $\overline{\mathbf{B}_o}$: (B, M, E, N) ← $\mathbf{\Delta}_o \otimes \mathbf{B}_o$
  /* initialize h_o and $\mathbf{y}_o$ with 0 */
  h_o : (B, E, N) ← zeros (B, E, N)
  $\mathbf{y}_o$: (B, M, E) ← zeros (B, M, E)
  /* SSM recurrent */
  for i in {0, ..., M-1} **do**
    $h_o = \overline{\mathbf{A}_o}[:, i, :, :] \odot h_o + \overline{\mathbf{B}_o}[:, i, :, :] \odot \mathbf{x}'_o[:, i, :, \mathbf{None}]$
    $\mathbf{y}_o[:, i, :] = h_o \otimes \mathbf{C}_o[:, i, :]$
  end for
end for
/* get gated $\mathbf{y}$ */
$\mathbf{y}'_{forward}$: (B, M, E) ← $\mathbf{y}_{forward} \odot \mathbf{SiLU}(\mathbf{z})$
$\mathbf{y}'_{backward}$: (B, M, E) ← $\mathbf{y}_{backward} \odot \mathbf{SiLU}(\mathbf{z})$
/* residual connection */
$\mathbf{T}_l$: (B, M, D) ← $\mathbf{Linear}^{\mathbf{T}}(\mathbf{y}'_{forward} + \mathbf{y}'_{backward}) + \mathbf{T}_{l-1}$
Return: $\mathbf{T}_l$

3．Vim 模块

原始的 Mamba 模块是为一维序列设计的，不适用于需要空间感知理解的视觉任务。在本节中，我们介绍 Vision Mamba 编码的基本构建模块 Vim 模块，如图 1 右侧所示。具体来说，像我们在算法 1 中所展示的操作。输入的标记序列T_{l-1}首先通过归一化层进行归一化。接下来，我们将归一化后的序列线性投影到维度大小为E的x和z。然后，我们从前向和后向两个方向处理x。对于每个方向，我们首先对x进行一维卷积，得到x'_o。

然后，我们将x_o'线性投影到B_o, C_o, Δ_o。Δ_o然后用于分别离散化得到$\bar{A}_o$和$\bar{B}_o$。最后我们通过 SSM 计算前向输出$y_{forward}$和反向输出$y_{backward}$，并通过z进行门控，并加在一起得到输出标记序列T_l。

4．效率优化

Vim 通过借助于 Mamba 的硬件友好的实现方式确保运行的效率。优化的关键思想是避免 GPU 的 I/O 瓶颈和内存瓶颈。

IO 高效性。高带宽存储器（HBM）和 SRAM 是 GPU 的两个重要组成部分。其中，SRAM 具有更大的带宽，而 HBM 具有更大的存储容量。标准的 Vim 的 SSM 操作在 HBM 上需要的 I/O 数量是 O(BMEN)，其中 B 为批量大小，M 为图像块序列长度，E 表示扩展状态维度，N 表示 SSM 维度。受到 Mamba 的启发，Vim 首先将 O(BME+EN)字节的内存从较慢的 HBM 读取到较快的 SRAM 中。然后 Vim 在 SRAM 中获取对应的参数，并执行 SSM 操作，最终将输出结果写回 HBM。此方法可以将 I/O 数量从 O(BMEN)降低到 O(BME+EN)从而大幅度提升效率。

内存高效性。为了避免内存不足问题并在处理长序列时降低内存使用，Vim 选择了与 Mamba 相同的重计算方法。对于尺寸为 (B.M,E,N)的中间状态的梯度，Vim 在网络的反向传递中重新计算它们。对于激活函数和卷积的中间激活值，Vim 也重新计算它们，以优化 GPU 的内存需求，因为激活值占用了大量内存，但重新计算速度很快。

计算高效性。Vim 模块中的 SSM 算法和 Transformer 中的自注意力机制都在自适应地提供全局上下文方面起到了关键作用。给定一个视觉序列$T \in \mathbb{R}^{1\times M\times D}$和默认的设置$E = 2D$。全局注意力机制和 SSM 的计算复杂度分别为：

$$\Omega(SelfAttention) = 4MD^2 + 2M^2D$$

$$\Omega(SSM) = 3M(2D)N + M(2D)N$$

其中，自注意力机制的计算复杂度和序列长度M成平方关系，而 SSM 的计算复杂度和序列长度M呈线性关系。这种计算效率使得 Vim 在处理具有长序列长度的千兆像素级别应用时具有良好的扩展性。

三、实验结果

该方法在标准的大型图片分类数据集ImageNet-1K上进行验证。并将分类训练好的模型作为预加载权重用于下游图片密集型预测任务中去，如COCO数据集上的目标检测和实例分割任务，ADE20K上的像素级别的语义分割任务。

1．分类对比

如表 1 与当前主流的分类模型对比 Vim 显示出了相当的精度，将 Vim 和基于 CNN、Transformer 和 SSM 的主干网络进行比较，Vim 显示了相当甚至更优的性能。例如，在参数量相同的情况下 Vim-Small 的准确率 80.3%，比 ResNet50[12]高出了 4.1 个百分点。与传统的基于自注意力机制的 ViT[4]相比，Vim 在参数数量和准确率上均有显著提升。与视觉 Transformer ViT 高度优化的变种 DeiT 相比，Vim 在不同模型尺度上均以相似的参数数量取得了更好的精度。

如表 1 所示，Vim 的优越的效率足以支持更细粒度

Method	image size	#param.	ImageNet top-1 acc.
Convnets			
ResNet-18	224^2	12M	69.8
ResNet-50	224^2	25M	76.2
ResNet-101	224^2	45M	77.4
ResNet-152	224^2	60M	78.3
ResNeXt50-32×4d	224^2	25M	77.6
RegNetY-4GF	224^2	21M	80.0
Transformers			
ViT-B/16	384^2	86M	77.9
ViT-L/16	384^2	307M	76.5
DeiT-Ti	224^2	6M	72.2
DeiT-S	224^2	22M	79.8
DeiT-B	224^2	86M	81.8
SSMs			
S4ND-ViT-B	224^2	89M	80.4
Vim-Ti	224^2	7M	76.1
Vim-Ti†	224^2	7M	78.3 +2.2
Vim-S	224^2	26M	80.3
Vim-S†	224^2	26M	81.4 +1.1
Vim-B	224^2	98M	81.9
Vim-B†	224^2	98M	83.2 +1.3

表 1　ImageNet-1K 分类骨干网络对比

的微调，在通过细粒度微调后，与基于 SSM 的 S4ND-ViT-B[13]相比，Vim 在参数数量减小 3 倍的情况下达到了相似的精度，Vim-Ti+，Vim-S+和 Vim-B+的结果均有所提高。其中，Vim-S+甚至达到了与 DeiT-B 相似的效果。

2．语义分割对比

在 ADE20K 语义分割数据集上，我们将 ImageNet-1K 上训练好的权重加载到 UperNet[14]分割器中，使用 Vim 作为骨干网络进行特征提取，如表 2 所示，Vim 取得了相比于 CNN 网络 ResNet 更少的参数量以及更高的精度，与去 Transformer 模型 DeiT 相比，Vim 取得了更优的精度。

3．目标检测与实例分割对比

在 COCO 目标检测与实例分割数据集上，我们将 ImageNet-1K 上训练好的权重加载到 Cascade-RCNN 框架中，使用 Vim 作为骨干网络进行特征提取，如表 3 所示，Vim 取得相对于 Transformer 的 DeiT 更好的检测框精度和实例分割精度。值得注意的是，在高清图像输入的目标检测任务上，图像输入分辨率为 1024×1024，由于 Transformer 的平方复杂度，需要将自注意力机制限制在固定大小的窗口内， 而 Vim 得益于其线性复杂度，无需窗口化，可以进行全局的视觉特征感知，从而取得了相对于表 3 中窗口化 DeiT 更好的精度。

Method	Backbone	image size	#param.	*val* mIoU
DeepLab v3+	ResNet-101	512^2	63M	44.1
UperNet	ResNet-50	512^2	67M	41.2
UperNet	ResNet-101	512^2	86M	44.9
UperNet	DeiT-Ti	512^2	11M	39.2
UperNet	DeiT-S	512^2	43M	44.0
UperNet	Vim-Ti	512^2	13M	41.0
UperNet	Vim-S	512^2	46M	44.9

表 2　ADE20K 语义分割对比

Backbone	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{box}_{s}	AP^{box}_{m}	AP^{box}_{l}
DeiT-Ti	44.4	63.0	47.8	26.1	47.4	61.8
Vim-Ti	45.7	63.9	49.6	26.1	49.0	63.2
Backbone	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{s}	AP^{mask}_{m}	AP^{mask}_{l}
DeiT-Ti	38.1	59.9	40.5	18.1	40.5	58.4
Vim-Ti	39.2	60.9	41.7	18.2	41.8	60.2

表 3　COCO 目标检测和实例分割对比

Bidirectional strategy	ImageNet top-1 acc.	ADE20K mIoU
None	73.2	32.3
Bidirectional Layer	70.9	33.6
Bidirectional SSM	72.8	33.2
Bidirectional SSM + Conv1d	73.9	35.9

表 4　双向 SSM 建模消融实验

Classification strategy	ImageNet top-1 acc.
Mean pool	73.9
Max pool	73.4
Head class token	75.2
Double class token	74.3
Middle class token	76.1

表 5　分类策略消融实验

4．消融实验

双向 SSM。如表 4 所示，双向 SSM 相较于原本的单向 SSM 取得了更高的分类精度， 且在下游的密集型预测任务上取得更为显著的优势。这一结果显示了本文提出的双向设计对于视觉特征学习的必要性与重要性。

分类策略。在表 5 中，我们探索了以下几种分类策略：

- *Mean pool*，将最后 Vision Mamba 编码器输出的特征进行平均池化。
- *Max pool*，将最后 Vision Mamba 编码器输出的特征进行最大化池化。
- *Head class token*，将类别标记词元置于图像块序列头部。
- *Double class token*，将类别标记词元置于图像块序列两端。
- *Middle class token*，将类别标记词元置于图像块序列中间。

如表 5 所示，实验结果表明，中间类别标记策略能够充分利用 SSM 的循环特性和 ImageNet 中的中心对象先验，展示了最佳的 top-1 准确率 76.1。

四、总结

该论文提出了 Vision Mamba (Vim)，以探索最新

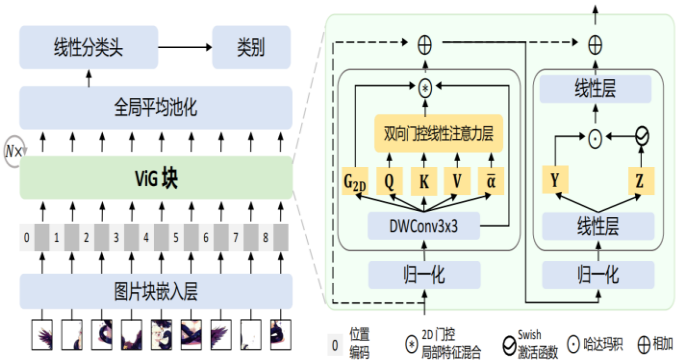


图 3　ViG 示意图

的高效状态空间模型 Mamba 作为通用视觉主干网络。与以往用于视觉任务的状态空间模型采用混合架构或等效的全局二维卷积核不同，Vim 以序列建模的方式学习视觉表示，并未引入图像特定的归纳偏置。得益于所提出的双向状态空间建模，Vim 实现了数据依赖的全局视觉上下文，并具备与 Transformer 相同的建模能力，同时计算复杂度更低。受益于 Mamba 的硬件感知设计，Vim 在处理高分辨率图像时的推理速度和内存使用显著优于 ViTs。在标准计算机视觉基准上的实验结果验证了 Vim 的建模能力和高效性，表明 Vim 具有成为下一代视觉主干网络的巨大潜力。

五、后续系列工作

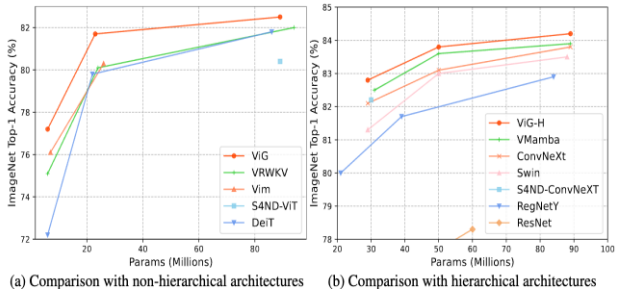
在Vision Mamba发表之后，华中科技大学团队围绕着高效线性的视觉表征学习进行了更进一步的探索，包括：ViG[16](Linear-complexity Visual Sequence Learning with Gated Linear Attention)，提出了基于线性门控注意力机制的视觉序列表征学习模型和DiG[17](Scalable and Efficient Diffusion Models with Gated Linear Attention)，将线性门控注意力单元引入图像生成领域，进行高效的图像生成。

1．图像理解骨干网络 ViG

尽管 Vision Mamba 这种线性复杂度序列建模网络在各种计算机视觉任务上取得了与视觉 Transformer 相似的建模能力，同时使用了更少的 FLOPs 和内存。然而，它们在实际运行速度方面的优势并不显著。

为了解决这个问题，我们引入了用于视觉任务的门控线性注意力（Gated Linear Attention, GLA），利用其卓越的硬件感知和效率。如图 3 所示，我们提出了方向性门控，通过双向建模捕获一维全局上下文，并通过二维门控局部注入自适应地将二维局部细节注入一维全局上下文。我们的硬件感知实现进一步将前向和后向扫描合并到一个单一的内核中，增强了并行性，降低了内存成本和延迟。

如图 4 和图 5 所示，ViG 模型在 ImageNet 及下游任务中在准确率、参数和 FLOPs 方面提供了有利的权衡，性能优于流行的 Transformer 和基于 CNN 的模型。实际的运行速度也是显著优于 Vision Mamba，能够和经过高度优化后的 Transformer 和 CNN 变体在低清图像输入时相当。值得注意的是，ViG-S 在仅使用 DeiT-B 27%的参数和 20%的 FLOPs 情况下，其准确率与 DeiT-B 相当，并且在 224×224 图像上的运行速度快 2 倍。在 1024×1024 分辨率下，ViG-T 使用的 FLOPs 少 5.2 倍，节省了 90%的 GPU 内存，运行速度快 4.8 倍，



(a) Comparison with non-hierarchical architectures　(b) Comparison with hierarchical architectures

图 4　ImageNet-1K 上和主流的非层级化结构 (a)层级化结构 (b)进行参数量和精度的平衡的比较

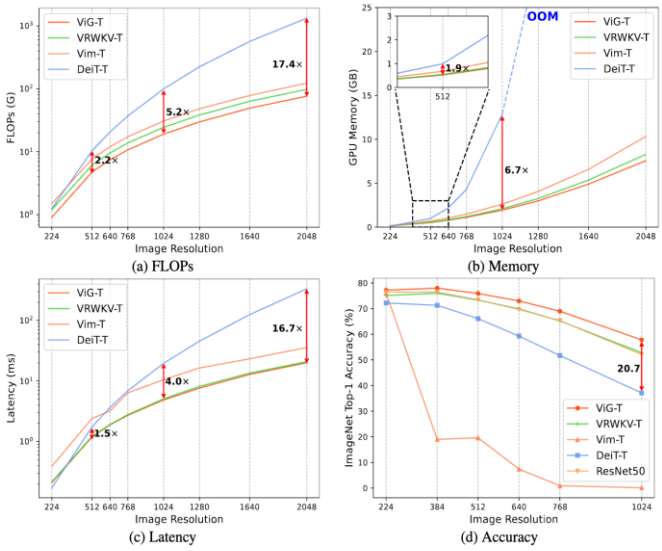


图 5　在 ImageNet-1K 上进行随着分辨率增大的高效性和精度对比

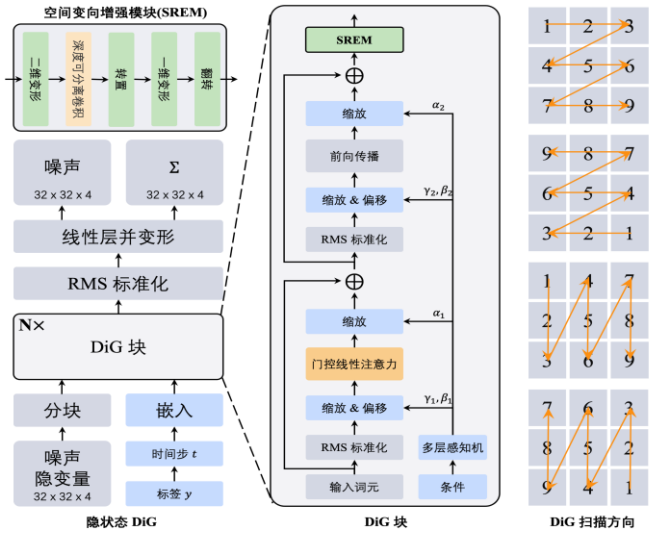


图 6　DiG 示意图

top-1 准确率比 DeiT-T 高 20.7%。这些结果表明，ViG 是一种高效且可扩展的视觉表示学习解决方案。

2．图像生成扩散模型 DiG

通过大规模预训练的扩散模型在视觉内容生成领域取得了显著成功， 其中 Diffusion Transformers (DiT) 尤为突出。然而，DiT 模型在可扩展性和二次复杂度效率方面面临挑战。这些挑战限制了 DiT 模型在处理高分辨率图像和长序列建模时的性能和效率。

为了解决这些问题，我们提出了 Diffusion Gated Linear Attention Transformers (DiG)，这是一种简单、易于采用的解决方案，具有最小的参数开销。DiG 遵循

Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
Previous state-of-the-art diffusion methods.					
ADM [13]	10.94	6.02	100.98	0.69	0.63
ADM-U	7.49	**5.13**	127.49	0.72	0.63
ADM-G	4.59	5.25	186.70	0.82	0.52
ADM-G, ADM-U	3.94	6.14	215.84	0.83	0.53
CDM [21]	4.88	-	158.71	-	-
LDM-8 [46]	15.51	-	79.03	0.65	0.63
LDM-8-G	7.76	-	209.52	0.84	0.35
LDM-4-G (cfg=1.25)	3.95	-	178.22	0.81	0.55
LDM-4-G (cfg=1.50)	3.60	-	247.67	**0.87**	0.48
Baselines and Ours.					
DiT-S/2-400K [39]	68.40	-	-	-	-
DiG-S/2-400K	62.06	11.77	22.81	0.39	0.56
DiT-B/2-400K	43.47	-	-	-	-
DiG-B/2-400K	39.50	8.50	37.21	0.51	0.63
DiT-L/2-400K	23.33	-	-	-	-
DiG-L/2-400K	22.90	6.91	59.87	0.60	0.64
DiT-XL/2-400K	19.47	-	-	-	-
DiG-XL/2-400K	18.53	6.06	68.53	0.63	0.64
DiG-XL/2-1200K	11.96	7.39	106.65	0.65	**0.67**
DiG-XL/2-1200K (cfg=1.5)	**2.84**	5.47	**250.36**	0.82	0.56

表 6　ImageNet-1K 图像生成精度对比

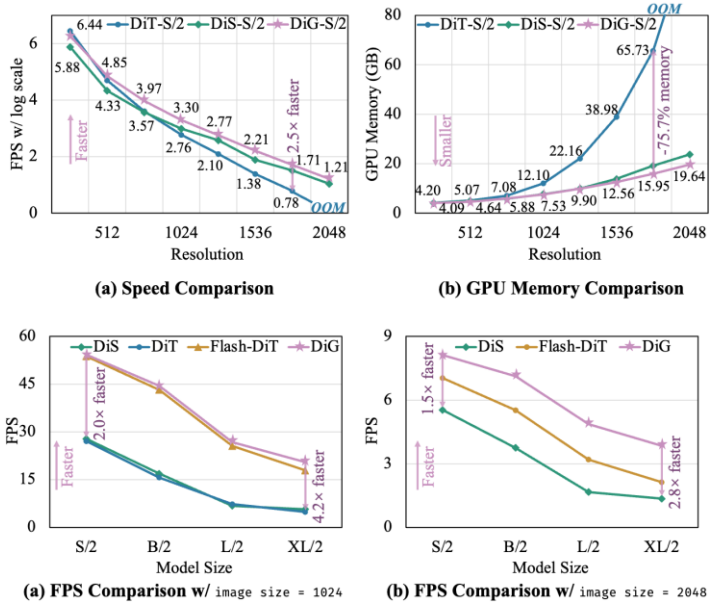


图 7　高效性对比

DiT 的设计，但在效率和有效性上更胜一筹。DiG 利用门控线性注意力（Gated Linear Attention，GLA）Transformer 的长序列建模能力，将其应用扩展到扩散模型。

DiG 的流程图如图 6 所示，它通过空间变向加强模块进行序列方向的控制并完成二维感知。与 DiT 模型的同期对比结果也更好。

如表 6 和图 7 所示，DiG 在 ImageNet 图像生成任务上精度和效率均表现优异。DiG-S/2 在图像尺寸为 1792*1792 时速度是 DiT-S/2 的 2.5 倍，节省了 75.7% 的 GPU 显存。同时 DiG 在更大的模型尺寸下依然保持了高效的特点。在图像尺寸为 1024 的情况下，DiG-XL/2 速度是 DiS-XL/2 的 4.2 倍；在输入图像大小为 2048 的情况下，DiG-XL/2 的速度是 CUDA 优化的 FlashAttention-2 加持下的 DiT 的 1.8 倍。

最后，本文介绍的 Vim、ViG、DiG 三个模型的代码和预训练模型均发布于：https://github.com/hustvl。

责任编委　魏秀参

参考文献

[1] Zhu L, Liao B, Zhang Q, et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. In ICML 2024.
[2] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database. In CVPR 2009.
[3] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context. In ECCV 2014.
[4] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR 2021.
[5] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention. In ICML 2021.
[6] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In ICCV 2021.
[7] Li Y, Mao H, Girshick R, et al. Exploring plain vision transformer backbones for object detection. In ECCV 2022.
[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In NeurIPS 2017.
[9] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. In Proceedings of the IEEE 1998.
[10] Gu A, Goel K, Ré C. Efficiently modeling long sequences with structured state spaces. In ICLR 2022.
[11] Gu A, Dao T. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.
[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In CVPR 2016.
[13] Nguyen E, Goel K, Gu A, et al. S4nd: Modeling images and videos as multidimensional signals with state spaces. In NeurIPS 2022.
[14] Xiao T, Liu Y, Zhou B, et al. Unified perceptual parsing for scene understanding. In ECCV 2018.
[15] Cai Z, Vasconcelos N. Cascade r-cnn: Delving into high quality object detection. In CVPR 2018.
[16] Liao B, Wang X, Zhu L, et al. ViG: Linear-complexity Visual Sequence Learning with Gated Linear Attention. arXiv preprint arXiv:2405.18425, 2024.
[17] Zhu L, Huang Z, Liao B, et al. DiG: Scalable and Efficient Diffusion Models with Gated Linear Attention. arXiv preprint arXiv:2405.18428, 2024.
[18] Peebles W, Xie S. Scalable diffusion models with transformers. In CVPR 2023.

朱良辉

华中科技大学电子信息与通信学院 2023 级博士研究生，导师为王兴刚教授，主要研究方向为生成模型。
Email:lhzh@hust.edu.cn

张 骞

张骞，博士毕业于中国科学院自动化研究所模式识别与智能系统专业，长期专注计算机视觉、机器学习、模式识别等技术方向的研究，发表论文 40 余篇。现为地平线算法研发总监，从事高阶自动驾驶技术在人工智能芯片上的探索创新与技术落地。获得了第十届吴文俊人工智能专项奖芯片项目一等奖、2022 年中国汽车工程学会“科学技术奖科技进步一等奖”。

Email: qian01.zhang@horizon.ai

廖本成

华中科技大学人工智能与自动化学院 2022 级博士研究生，导师为王兴刚教授，主要研究方向为目标检测与感知，自动驾驶。

Email: bcliao@hust.edu.cn

王鑫龙

王鑫龙， 智源研究院视觉模型研究中心负责人。本科毕业于同济大学，博士毕业于澳大利亚阿德莱德大学，师从沈春华教授。他的研究兴趣是计算机视觉和基础模型，近几年研究工作包括视觉感知 （SOLO， SOLOv2），视觉表征 (DenseCL， EVA)，视觉上下文学习(Painter，SegGPT)，多模态表征(EVA-CLIP，Uni3D)，多模态上下文学习(Emu，Emu2)。入选 Google PhD Fellowship、国家海外高层次青年人才。

Email: wangxinlong@baai.ac.cn

刘文予

刘文予，1963 年 9 月生，1986 年本科毕业于清华大学计算机系，1991 年硕士毕业于华中理工大学电信系，2001 年博士毕业于华中科技大学电信系，之后一直于华中科技大学电信学院工作，目前任电信学院二级教授。中国通信学会会士、中国图象图形学会图象视频通信专业委员会主任委员、中国图象图形学会常务理事、中国电子教育学会研究生教育分会常务理事。

Email: liuwy@hust.edu.cn

王兴刚

王兴刚，电信学院教授、博导，现任 Elsevier Image and Vision Computing 期刊共同主编，入选国家青年人才计划，获湖北省青年五四奖章获，CSIG 青年科学家奖，吴文俊人工智能优秀青年奖，CVMJ 最佳论文奖，互联网+金奖，华为/vivo 等公司优秀技术合作项目奖。 分别与 2009 年和 2014 年在华中科技大学获得了学士和博士学位，美国 UCLA 和天普大学访问学者。主要研究方向为基座模型、视觉表征学习、目标检测分割跟踪，在顶级期刊会议上发表论文 60 余篇，引用超过 2.8 万次。

Email: xgwang@hust.edu.cn

热点追踪

基于深度学习的深度图像填充

西北工业大学 王宇飞 戴玉超

一、引言

获取准确且稠密的场景深度在自动驾驶和增强现实等多种应用中起着至关重要的作用[1]。然而，现有的深度传感器都无法捕获完整的场景深度。例如，自动驾驶汽车中常用的激光雷达获取的深度十分稀疏，通常无法直接使用。因此，深度填充任务[1][2]，即从深度传感器获取的稀疏深度图中估计稠密的深度图，已经吸引了工业界和研究界的广泛研究兴趣。

最近，基于深度学习的深度图填充方法通过大量堆叠的滤波器将稀疏深度图直接映射为稠密深度图，已经取得了优异的性能[3]。由于 RGB 图像包含丰富的语义信息，这些信息对于填充未知深度至关重要，因此 RGB 信息通常被用于指导深度填充[4]。尽管现有工作已经提出了许多先进的网络[5][6]，但从稀疏输入深度图和相应的 RGB 图像中直接预测准确且稠密的深度图仍然困难。现有方法通常需要使用数千万可学习参数学习鲁棒特征。例如，PENet[7]中包含了 132M 参数。如此大规模的网络通常需要大量计算资源，在现实场景中难以应用，而如果简单减少网络规模，方法性能会显著下降。此外，通过直接回归获得的预测深度图会出现模糊效果和物体边界的失真，需要通过额外的优化模块进一步优化。例如，流行的空间传播网络（Spatial Propagation Network，SPNs）通过递归操作更新直接回归方法的输出。因此，设计一个在保持效率的同时表现更好的高效深度填充架构对于进一步研究至关重要。

本文提出了一种用于深度填充的新型轻量级深度网络框架，即长短范围循环更新网络（Long-short Range Recurrent Updating，LRRU）。与现有的直接回归方法不同，LRRU 通过迭代更新由非学习方法获得的初始深度图得到精确深度图。初始深度图具有粗略但完整的场景深度信息，可以帮助减轻网络直接从稀疏输入深度图回归精确稠密深度的学习负担。尽管现有的空间传播网络（SPNs）已经表明，深度图可以通过学习的建模每个像素的相关点及其亲和性的空间变化内核来优化，但这些方法存在以下限制无法直接用于所提框架：(a) 内容无关的更新单元：更新所需的内核参数由 RGB 和稀疏深度的特征预测，不能自适应地根据目标深度图（即将被更新的深度图）调整；(b) 不灵活的循环策略：在更新过程中，内核范围是固定的，需要多次迭代才能获得长距离依赖和令人满意的结果。

为了解决上述问题，本文提出了目标依赖更新（Target-dependent Update Unit ，TDU）单元和长短范围循环策略，使迭代更新过程具备内容自适应性和高度灵活性。TDU 通过结合交叉引导和自引导特征来预测相关点的采样位置及其与参考点之间的权重（亲和性）。从 RGB 图像和稀疏深度中提取的交叉引导特征可以指导 TDU 避免无关的相关点，而从将要更新的深度图中提取的自引导特征使 TDU 能够适应目标深度图的内容。此外，TDU 还通过学习残差进一步提高了性能。观察发现，当更新过程中的多个 TDU 分别采用不同尺度的交叉引导特征时，较小尺度交叉引导特征指导的 TDU 会自适应地学习在较大范围内获取相关点，反之亦然。由于本文的初始深度图是通过扩展稀疏测量点获得的，因此大多数像素的周围点是不准确的。因此，在更新过程的初始阶段，本文使用小尺度交叉引导特征引导 TDU 预测一个大范围，以便获取一些远距离但准确的点作为相关点。随着深度图变得更加精细，逐步使用较大

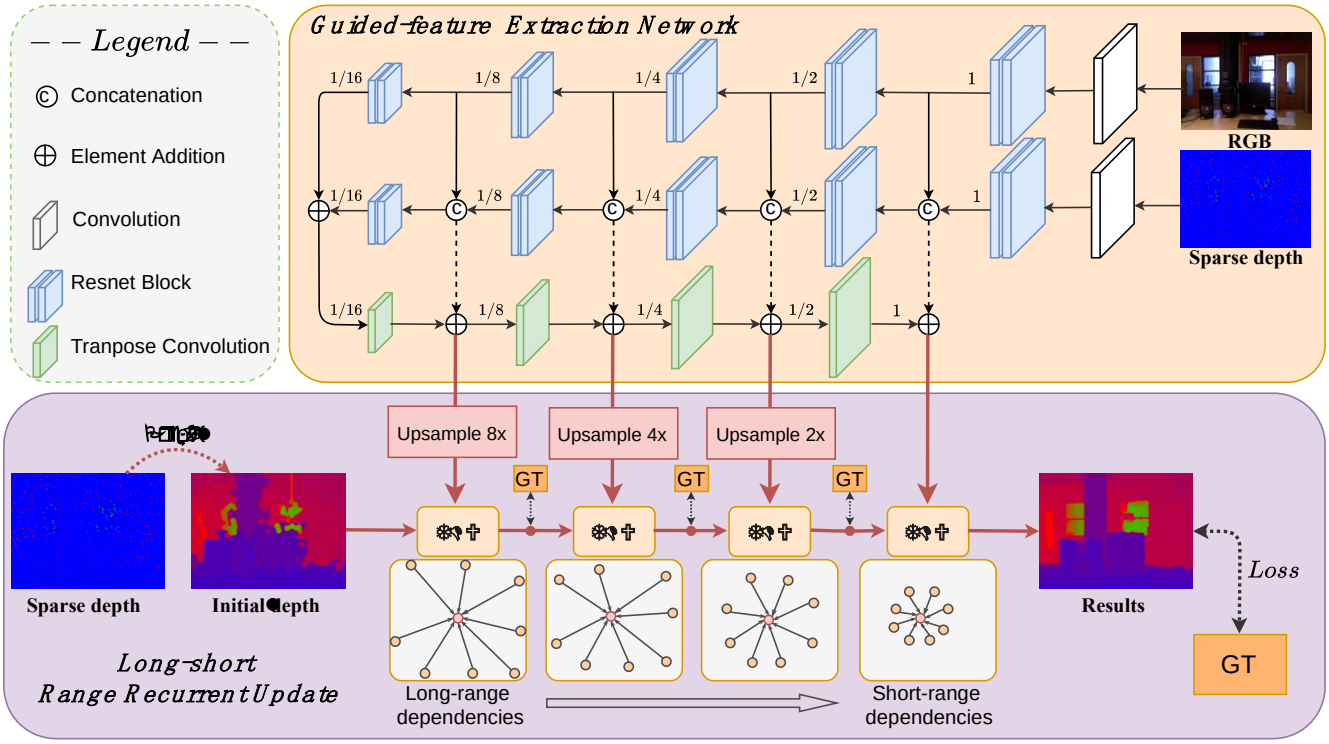


图 1 方法框架。长短范围循环更新网络（LRRU）从 RGB 图像和稀疏深度图中提取交叉引导特征，然后根据所提出的长短范围循环策略，通过目标依赖更新模块（TDU）迭代更新预填充的深度图。

尺度的交叉引导特征以更多地关注短距离相关点。

在室外数据集[8]和室内数据集[9]上的大量实验结果验证了本文方法的性能。实验结果显示 LRRU 在不同参数范围内均达到了最先进的性能。特别是，LRRU-Base 模型在 NYUv2 和 KITTI 基准测试上的性能优于 SOTA 方法，并在 KITTI 榜单上排名第一。

二、深度图填充相关工作

在深度学习时代，最简单直接的深度填充方法采用直接回归的方法，使用各种网络结构从稀疏深度图预测稠密深度图。由于 RGB 图像包含丰富的纹理和语义信息，这些信息对于恢复深度图的结构细节至关重要，许多流行的方法融合了 RGB 图像和稀疏深度图的信息以增强深度填充。Ma 等人[3]提出将深度图和 RGB 图像连接形成 4D 张量，这被称为“早期融合”。为了缩小不同模态之间的差距，一些工作提出了“后期融合”方法[4][10][11][12]，分别提取 RGB 图像和稀疏深度图的特征，然后将融合特征输入网络。此外，Tang 等人[13]提出了一种基于引导动态卷积网络的特征融合模块，以更好地利用 RGB 图像的引导特征。此外，一些工作[14][15]提出先通过经典方法对稀疏深度进行稠密化，然后学习初始深度近似的残差。然而，现有方法通常使用大量参数和计算资源以获得良好结果，缺乏轻量且高效的网络架构。

直接回归方法预测的深度图存在模糊效果和物体边界失真的问题。为了解决这个问题，一些研究提出了一系列空间传播网络（SPNs），通过聚合参考像素和邻近像素来迭代更新直接回归方法的输出。最初的 SPN[16]通过上一行或列的三个相邻像素更新每个像素。串行更新过程分别在四个方向上进行，并通过最大池化组合结果。为了使更新过程更高效，Cheng 等人[17]提出了卷积空间传播网络（CSPN），它在固定局部邻域内同时更新所有像素。然而，固定局部邻域配置会引入无关点。此外，CSPN++[18]通过使用不同的核大小来组合结果。DSPN[19]和 NLSPN[20]通过学习到规则网格的偏移来预测非局部邻域，DSPN 通过计算特征之间的相似性获得核权重，而 NLSPN 通过网络学习获得。尽管改进的基于 SPN 的方法在选择相关点方面提供了更多的灵活性，但它们在更新过程中使用固定的核权重，限制了 SPN 的表示能力。为了解决这个问题，DySPN[21]通过学习到的注意力图给予不同距离的相关点可变权重。GraphCSPN[22]利用图神经网络（GNN）将 3D 信息集

成到更新过程中。然而，现有的 SPNs 仍然使用固定相关点，无法在更新过程中动态调整它们。

三、深度图像填充方法

给定一个稀疏深度图，所提方法首先通过一种简单的非学习方法[23]对其进行稠密化处理。然后，根据长短范围循环策略，通过目标依赖更新模块迭代更新初始深度图，以获得准确且稠密的深度图。为了方便描述，本文使用“目标深度”来指代在第t次更新中要更新的深度图。因此，$\hat{D}^{1}$表示通过上述非学习方法获得的初始深度图。此外，$\hat{D}^{t+1}$表示更新后的结果。

1．目标依赖更新模块

目标依赖更新模块（TDU）通过学习到的空间可变内核更新目标深度图，这些核建模了每个像素的相关点及其关联性。为了避免固定局部邻域配置带来的无关相关点，TDU 采用全卷积网络来预测核权重和相关点的浮点采样位置，其中浮点采样位置通过学习到规则网格的偏移量来获得。然而，通常情况下，核权重和偏移量的直接监督信息是不可用的，通常会导致训练不稳定。

为了解决上述问题，本文利用来自 RGB 图像和稀疏深度图的特征来引导 TDU 以获得相关点。由于稠密的 RGB 图像和稀疏的深度图属于不同的模态，本文采用了类似于双编码器网络的方法，该方法使用两个独立的子网络分别提取 RGB 图像和稀疏深度图的特征，并在多个尺度上融合这些特征。本文将来自 RGB 图像和稀疏深度图的特征称为交叉引导特征，而将来自目标深度图的特征称为自引导特征。如下面公式所示，交叉引导特征$F_{\text{Cross-guided}}$通过特征提取网络f_{θ}从输入的 RGB 图像I和稀疏深度图S中提取，而自引导特征$F_{\text{Self-guided}}$则通过卷积层f_{ψ}从目标深度图$\hat{D}^{t}$中获取。

$$F_{\text{Cross-guided}} = f_{\theta}(I,S), F_{\text{Self-guided}} = f_{\psi}\left(\hat{D}^{t}\right) \quad (1)$$

具体而言，TDU 首先将交叉引导特征和自引导特征拼接起来，然后通过两个独立的 1 x 1 卷积层学习权重和偏移特征图。为了使权重和偏移快速收敛，本文对它们的行为进行了限制并指导学习过程。权重特征图有k^{2}个通道，其中k是核大小，在本文中设为 3。本文使用一个层使权重大于零且小于一。此外，本文从层的输出中减去平均值，使权重的和为零，这样的操作类似于高通滤波器。偏移特征图有$2k^{2}$个通道，表示在 x 和 y 方向上采样点相对于规则网格位置的偏移量。然而，为了确保每个参考像素参与其自身的更新过程，首先预测一个具有$2(k^{2}-1)$个通道的偏移特征图，然后在偏移特征图通道的中心插入零。

由于更新单元的输入和输出图高度相关，并共享低频信息，本文提出学习目标深度图的残差图像，以增强结构细节和抑制噪声。给定学习到的权重和采样偏移，如下公式所示，位置$\mathbf{p}=(x,y)$处的残差图像通过加权平均获得。

$$\Delta\hat{D}_{\mathbf{p}}^{t} = \sum_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \mathbf{W}_{\mathbf{pq}}\left(F_{\text{Cross-guided}}, F_{\text{Self-guided}}\right)\hat{D}_{\mathbf{q}}^{t} \quad (2)$$

在上述公式中，$\mathbf{q}\in\mathcal{N}(\mathbf{p})$ 表示位置$\mathbf{p}$的邻域集合，$\mathbf{W}_{\mathbf{pq}}\left(F_{\text{Cross-guided}}, F_{\text{Self-guided}}\right)$是位置$\mathbf{p}$的权重。由于偏移量通常是小数，本文使用双线性插值来采样局部的四个点。滤波器权重是从交叉引导特征和自引导特征中预测得到的。本文从稀疏选择的位置中聚合深度值并加权。然后，本文将残差图像添加到目标深度图上，如下公式所示，以获得更新后的深度图$\hat{D}^{t+1}$。

$$\hat{D}^{t+1} = \hat{D}^{t} + \Delta\hat{D}^{t} \quad (3)$$

2．长短范围循环策略

通过非学习方法获得的初始深度图，只有少量可用的稀疏测量点及其周围点的准确性较高，大多数像素的周围点是不准确的。因此，在更新过程的初始阶段，应采用较大的核范围，以获得一些长距离但准确的点作为相关点。随着深度图变得更加精细，核范围应逐渐缩小，以更多地关注与目标点相近的短距离点。

因此，本文提出了一种长短范围循环更新策略。每个 TDU 的参数，包括核权重和相关点的采样位置，都是通过考虑交叉引导和自引导特征来学习的。通过观察发现，当迭代更新过程中的 TDU 分别由不同尺度的交叉引导特征引导时，由较小尺度交叉引导特征引导的 TDU 将自适应地学习以获取较大范围的相关点，反之亦然。本文认为这是由于不同尺度的交叉引导特征具有不同的感受野。基于上述观察，本文首先采用 1/8 尺度的交叉引导特征图来引导第一轮迭代的 TDU，以获取较大

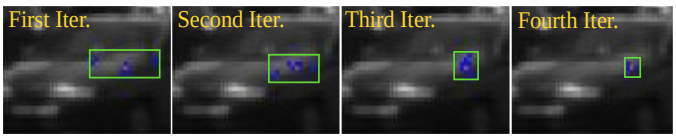


图 2 在迭代更新过程中由大到小的核范围是如何动态调整以捕捉长到短范围的依赖关系（红点表示参考像素，蓝点表示相关点）

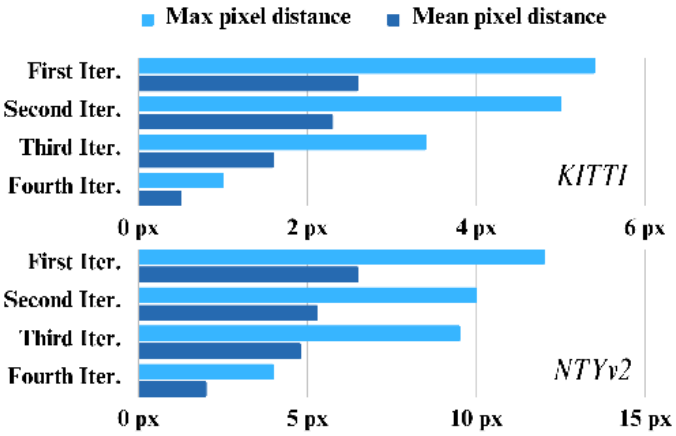


图 3 在 KITTI 和 NYUv2 数据集上分析相关点到参考像素的最大和平均像素距离

范围的相关点。在后续的迭代中，TDU 逐渐使用较大尺度的交叉引导特征，例如 1/4 尺度、1/2 尺度和全尺度，以获取较小范围的相关点。图 2 和图 3 中的示例展示了在 KITTI 和 NYUv2 数据集上的迭代更新过程中，内核范围从大到小的变化。由于所提出的循环更新策略具有高度灵活性，本文可以通过较少的迭代和相关点数量来实现令人满意的结果。

三、深度图像填充实验结果

为验证所提算法的有效性和先进性，本节在 KITTI

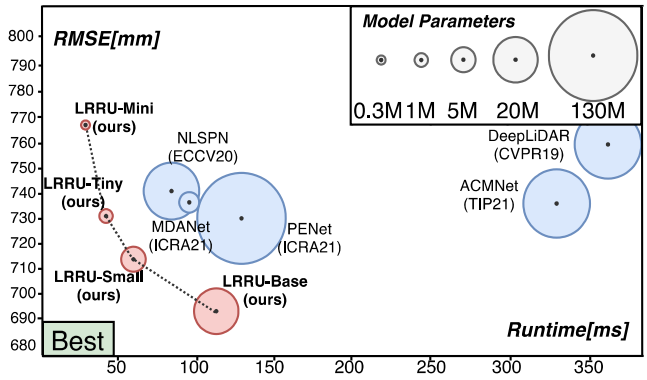


图 4 LRRU 在 KITTI 数据集在不同参数范围内均达到了最先进的性能

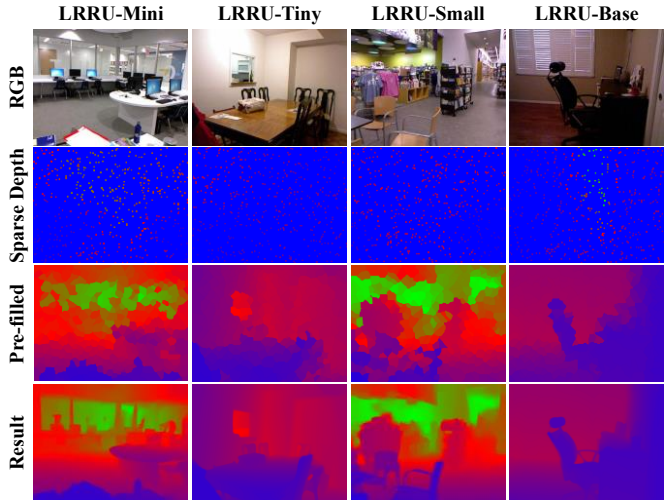


图 5 LRRU 在 NYUv2 数据集上的定性结果。稀疏深度图被扩展以显示，“预填充”表示预填充的深度图。由于稀疏深度图中可用的点很少，预填充的深度图非常粗糙。然而，LRRU 可以将其更新为令人满意的结果。

数据集和 NYUv2 数据集上进行了大量的实验。KITTI 数据集是一个流行的真实世界自动驾驶数据集，由从原始 LiDAR 扫描投影的稀疏深度图和对应的 RGB 图像组成。它包含 86,000 张图像用于训练，1,000 张图像用于验证，以及 1,000 张无真实值的测试图像，需要在 KITTI 在线基准测试上进行测试。NYUv2 数据集由从 464 个不同室内场景获取的 RGB 和深度图像组成。按照标准设置，本文使用从训练集中采样的 50,000 张图像训练模型，并在 654 张官方标注图像上进行测试。对于训练和测试数据集，原始大小为 640 × 480 的图像被下采样到一半，然后居中裁剪到 304 × 228。

如图 4 所示，在 KITTI 数据集上的实验结果显示了 LRRU 在保持效率的同时表现更好。图 5 中展示了 LRRU 在 NYUv2 数据集上的定性结果，由于 NYUv2 数据集的稀疏深度图稀疏程度较高，通过简单的手工方法获得的预填充深度图十分粗糙。然而，LRRU 通过循环更新过程可以修正初始深度图，得到了令人满意的结果，即使在一些细小结构中也能表现良好。

四、总结与展望

本文提出了一种长短范围循环更新网络的新型轻量级深度填充网络框架。基于所提出的目标依赖更新单元和长短范围循环策略，本文的迭代更新过程具有内容

自适应性和高度灵活性。与传统的直接回归方法相比，本文的方法在参数更少和推理时间更短的情况下实现了优越的性能。实验结果表明，所提方法在室内和室外场景中性能均优于现有方法。

责任编委 储珺

参考文献

[1] Hu J, Bao C, Ozay M, et al. Deep depth completion from extremely sparse data: A survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (7): 8244-8264.

[2] Xie Z, Yu X, Gao X, et al. Recent advances in conventional and deep learning-based depth completion: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(3): 3395-3415.

[3] Ma F, Karaman S. Sparse-to-dense: Depth prediction from sparse depth samples and a single image[C]. Proceedings of the IEEE International Conference on Robotics and Automation. 2018: 4796-4803.

[4] Ma F, Cavalheiro G V, Karaman S. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera[C]. Proceedings of the IEEE Inter- national Conference on Robotics and Automation. 2019: 3288-3295.

[5] Rho K, Ha J, Kim Y. Guideformer: Transformers for image guided depth completion[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2022: 6240-6249.

[6] Zhang Y, Guo X, Poggi M, et al. Completionformer: Depth completion with convolutions and vision transformers[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 18527-18536.

[7] Hu M, Wang S, Li B, et al. Penet: Towards precise and efficient image guided depth completion[C]. Proceedings of the IEEE International Conference on Robotics and Au- tomation. 2021: 13656-13662.

[8] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The Inter- national Journal of Robotics Research, 2013, 32(11): 1231-1237.

[9] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from RGBD images[C]. Proceedings of the European Conference on Computer Vision. 2012.

[10] Yan Z, Wang K, Li X, et al. Rignet: Repetitive image guided network for depth completion [C]. Proceedings of the European Conference on Computer Vision. 2022: 6240-6249.

[11] Qiu J, Cui Z, Zhang Y, et al. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 3313-3322.

[12] Zhou W, Yan X, Liao Y, et al. BEV@ DC: Bird's-Eye View Assisted Training for Depth Completion[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 9233-9242.

[13] Tang J, Tian F P, Feng W, et al. Learning guided convolutional network for depth com- pletion[J]. IEEE Transactions on Image Processing, 2020, 30: 1116-1129.

[14] Wong A, Fei X, Hong B W, et al. An adaptive framework for learning unsupervised depth completion[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 3120-3127.

[15] WongA,CicekS,SoattoS.Learningtopologyfromsyntheticdataforunsupervisedepth completion[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1495-1502.

[16] Liu S, De Mello S, Gu J, et al. Learning affinity via spatial propagation networks[J]. Advances in Neural Information Processing Systems, 2017, 2017: 1521-1531.

[17] Cheng X, Wang P, Yang R. Learning depth with convolutional spatial propagation net- work[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(10): 2361-2379.

[18] Cheng X, Wang P, Guan C, et al. Cspn++: Learning context and resource aware convo- lutional spatial propagation networks for depth completion[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 10615-10622.

[19] Xu Z, Yin H, Yao J. Deformable spatial propagation networks for depth completion[C]. Proceedings of the IEEE International Conference on Image Processing. IEEE, 2020: 913-917.

[20] Park J, Joo K, Hu Z, et al. Non-local spatial propagation network for depth completion [C]. Proceedings of the European Conference on Computer Vision. 2020: 120-136.

[21] Lin Y, Cheng T, Zhong Q, et al. Dynamic spatial propagation network for depth com- pletion[C]. Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 1638- 1646.

[22] Liu X, Shao X, Wang B, et al. Graphcspn: Geometry-aware depth completion via dynamic gcns[C]. Proceedings of the European Conference on Computer Vision. 2022: 90-107.

[23] Ku J, Harakeh A, Waslander S L. In defense of classical image processing: Fast depth completion on the cpu[C]. Proceedings of the Conference on Computer and Robot Vision. IEEE, 2018: 16-22.

王宇飞

西北工业大学电子信息学院博士生，研究方向：深度填充，深度估计。
Email: wangyufei1951@gmail.com

戴玉超

西北工业大学电子信息学院教授、博士生导师，国家级青年人才。主要研究工作集中在机器视觉、智能感知、图像处理、智能无人系统等领域。主持国家自然科学基金、科技部科技创新 2030“新一代人工智能”重大研究计划子课题、JKW 领域基金重点项目等科研项目。近年来在 TPAMI、IJCV、ICCV、CVPR、NeurIPS、ECCV 等国际顶级期刊和会议上发表论文 70 余篇，谷歌学术引用超过 11000 次，H 因子 49。获得 CVPR 2012 最佳论文奖（大陆高校 30 年来首次获得该奖项）、陕西省自然科学奖一等奖、中国图象图形学学会青年科学家奖、火箭军“智箭火眼”人工智能挑战赛全国第一名、CVPR 2020 最佳论文奖提名等奖项。担任 APSIPA 杰出讲者和 CVPR、ICCV、ECCV、NeurIPS 等国际顶级会议领域主席。
Email: daiyuchao@nwpu.edu.cn

热点追踪

可信冲突多视角学习算法

西安电子科技大学 徐偲 司佳俊 管子玉 赵伟

本文解读了西安电子科技大学团队发表在AAAI 2024并获得最佳论文奖的工作。多视角学习旨在通过联合来自不同视角的数据信息，有效地利用多视角数据的共性和特性，从而更精准地揭示事物内在的本质特征。随着多视角学习向更多应用领域不断拓展，越来越多的学习任务面临更为真实的复杂开放环境。相比于传统“实验室”数据环境，复杂开放环境中具有更多的未知性，其中重要一点就是收集到的多视角数据往往存在着视角信息冲突的情况。以往解决这类问题的主要策略包括删除或替换冲突视角，但实际应用需要对冲突样本进行可靠的决策。针对视角冲突的多视角数据，本文提出了一种新的可靠冲突多视角学习问题，该问题要求模型为冲突的多视角数据提供决策结果和可靠性。因此，本文提出了一种可信冲突多视角学习算法，旨在利用视角内的不确定性信息和视角间的一致性信息来提高模型的学习能力，并通过一个简单有效的平均池化层聚合冲突观点。最后，本文对该方法进行了分析和验证，证明了该方法可以准确地建模多视角公共可靠度和视角特定可靠度之间的内在联系。

一、引言

人工智能系统通常通过多视角数据来感知和理解世界。例如，自动驾驶汽车系统通过多个传感器（如摄像头、激光雷达）感知周围环境；推荐系统从用户的多视角生成的内容（如文本和图片）中捕获用户的偏好。整合多个视角的一致和互补信息可以获得更全面的数据描述，从而促进聚类[1-3]，检索[4]和推荐[5,6]等各种任务。以往关于多视角学习[7-9]的大部分研究总是假设不同视角的数据是严格对齐的。例如，在分类任务中，不同的视角始终属于基本事实类别。然而，在现实环境中，这种假设可能并不成立。图 1 可视化了用户多视角生成内容的一个案例：文本和图像表示的食品类别存在冲突。因此，这种不同视角信息的冲突使得大多数多视角学习方法不可避免地退化甚至失败。

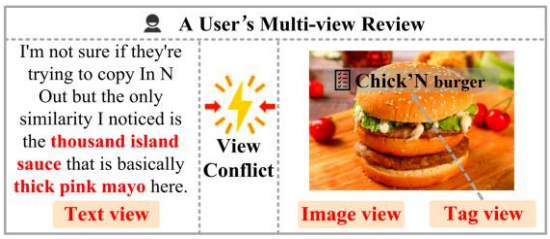


图 1　视角冲突的多视角数据

当前解决该问题的主要方法是消除冲突数据样本。最初的思路是将冲突数据视为异常值。该类方法通常包括 3 个步骤：1) 测量视角间的一致性；2) 识别异常值，即：不同视角之间存在着的显著不一致的样本；3) 去除异常值，构建干净的数据集。最近，一些多视角学习方法致力于学习原始数据的对齐关系，并相应地构建新的数据样本。例如，将图 1 中的文本将被替换为另一个对齐样本的视角，从而解决原样本中的冲突问题。

然而，现实世界的应用程序通常需要为冲突样本做出决策，而不仅仅是消除它们。例如，推荐系统需要从用户的冲突多视角评论中预测用户的偏好。考虑到冲突样本的决策可能是不可靠的，我们需要模型能够回答“该决策可靠吗?”因此，本文提出了一个新的问题——可靠的冲突多视角学习 (Reliable Conflictive Multi-view Learning，RCML) 问题，该问题要求模型对冲突多视角数据提供决策结果和附加的信度。

本文实现了一个可信冲突多视角学习算法——证据冲突多视角学习(Evidential Conflictive Multi view Learning，ECML)方法，整个网络框架如图 2 所示，它采用两个阶段来识别整合多视角信息，其中第一阶段以特定视角信息作为输入，采用证据神经网络来提取证据并得到特定视角的狄利克雷分布观点，第二阶段是利用冲突观点的聚合策略机制来整合多视角观点信息。本文的主要贡献总结如下：(1) 认识到在处理冲突的多视角数据时显式地提供决策结果和相关可靠性的重要性。(2) 提出了一种冲突观点的聚合策略和冲突度量，并从理论上证明了它可以准确地建模多视角共同可靠度和视角特定可靠度。(3) 实验结果验证了提出的模型在六个基准数据集上取得了领先的性能。这直接证明了该模型在冲突的多视角学习任务中的有效性和优越性，为该领域提供了更全面、更高效的解决方案。

二、相关工作

1．冲突的多视角学习

多视角学习通过整合多个互补视角的信息，能够有效地提升模型的泛化性能和适应性，但之前的研究大多假设不同视角的数据是严格对齐的。因此，在面对不同视角信息不一致的情况时，大多数现有的多视角学习方法不可避免地退化甚至失败。目前针对冲突的多视角学习问题，现有的方法主要分为两类：多视角异常点检测方法和部分视角对齐的多视角学习方法。

多视角异常点检测方法旨在检测特定环境中具有异常行为的异常点。Gao 等人[10]提出的水平异常检测算法是解决该问题的第一个方法，该算法构造一个集合相似矩阵并通过谱嵌入计算共识表示。另一种代表性的方法是由 Alvarez 等人[11]提出的亲和传播方法。它计算每个视角中每个样本与其相邻样本之间的差异来检测异常点。Li 等人[12]提出了一种多视角低秩分析框架，该框架执行交叉视角低秩分析，并采用精心设计的标准计算每个样本的异常值得分。Zhao 等人[13]提出了一种共识正则化离群值检测方法，该方法将所有视角的指示矩阵近似为具有共识正则化的共享矩阵以消除异常的影响。Li 等人[14]提出了一种能够从任意数量的数据视角中检测异常值的方法，该方法学习所有视角数据的潜在判别表示，并基于潜在判别表示定义了一个异常值评分函数。Hou 等人[15]提出了一种快速的低秩潜在表示用于大规模多视角异常点检测，该方法利用特定于视角的判别潜在表示来探索样本的判别成分，并利用深度编码器架构实现样本的异常分数。

在多视角学习领域，大多数现有工作的成功在很大程度上依赖于视角一致性的假设。换言之，一致性假设要求来自不同视角的数据必须严格对齐。然而，在实际操作中，数据很容易在收集或传输过程中出现错位，从而导致部分视角对齐问题。因此，人们提出了一些解决这一问题的传统方法。Lampert 等人[16]提出了弱配对

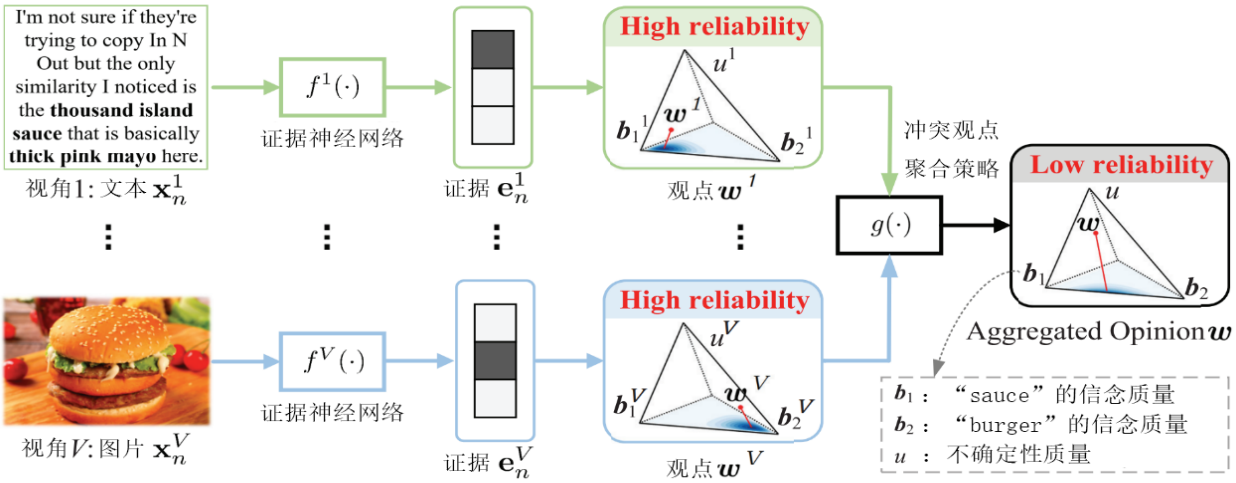


图 2　网络架构图。首先，利用特定视角的神经网络收集证据，即对每个类别的支持量。然后，该方法形成由所有类别的信念质量和不确定性组成的特定观点。最后，通过冲突观点聚合策略来整合观点。

最大协方差分析来克服弱配对数据的局限性。Zhang 等人[17]提出了少量的视角间约束代替映射来获取视角间的相互信息。目前，捕获数据高维特征的深度部分视角对齐方法尚处于起步阶段。Huang 等人[18]提出了一种称为部分视角对齐聚类的多视角聚类方法，该方法实现了匈牙利算法的可微代理来建立未对齐数据的对齐关系。Yang 等人[19]提出了使用噪声鲁棒对比损失同时学习数据高层表示和对齐数据的方法。Zhang 等人[20]提出了一种自适应视角对齐和特征增强网络，该网络使用自适应视角对齐模块和自聚焦机制来计算对齐矩阵。

现有的方法虽然可以解决部分的冲突多视角学习问题，但仍然存在着以下限制：1) 多视角异常点检测方法不能很好地检测出由模型本身引起的离群值，也不能给出一个直观的度量离群值程度的指标，并且现有的多视角异常点检测方法致力于学习原始数据的对齐关系，并据此构造新的数据样本，忽略了冲突样本的信息。2) 目前解决部分视角对齐问题的方法是存在问题的。现有的该类方法将其他视角对齐到一个视角上，这一操作改变了原始的数据分布，改变了数据集的整体结构和样本样本的真实信息。总之，以前的方法的目的是为了消除冲突的样本，而实际的应用程序通常需要为这些冲突的样本做出决策分析。

2．感知不确定性深度学习

深度神经网络已经在各种任务中取得了显著的成功，但其往往无法捕获其决策的不确定性，特别是对于低质量数据。不确定性可分为任意不确定性(数据不确定性)和认知不确定性(模型不确定性)。目前，不确定性量化的深度学习[21]可以分为:单一确定性方法[22,23]，贝叶斯方法[24-26]，集成方法[27]和测试时间增强方法[28]。证据深度学习(Evidential deep learning, EDL)[22]是单一确定性方法的代表性方法，其根据单个神经网络计算特定类别的证据。近年来，研究者将 EDL 扩展到多视角学习领域。开创性工作可信多视角分类(TMC)[29]利用 Dempster-Shafer 证据理论对不同视角的观点进行融合。随后，多种观点聚合方法[30-33]被研究者们提了出来。这些方法的一个重要特征是将另一种观点整合到原来的观点中，得到的不确定性质量会减小。但是，当面对现实中复杂的多视角数据时，尤其是在纳入不可靠或冲突的观点时，这些方法可能会出现错误。而且，本文认为最终融合后观点的不确定性应该与被融合的视角质量相关，而不是一味地减小。

三、网络框架

1．问题定义

假设多视角数据集为$\{\{\boldsymbol{x}_n^v\}_{v=1}^V, y_n\}_{n=1}^N$有$N=\bar{N}+\tilde{N}$个样本，每个样本有$V$个视角。其中有$\bar{N}$个正常样本，$\tilde{N}$个冲突样本，如图 3 所示。$\boldsymbol{y}_n \in \{0,1\}^K$是$y_n$对应的 one-hot 编码，$K$为类别数。RCML 的目标是为测试样本准确地预测类别y_n，并提供附加的预测不确定性$u_n \in [0,1]$[1]来度量决策可靠性$1-u_n$。

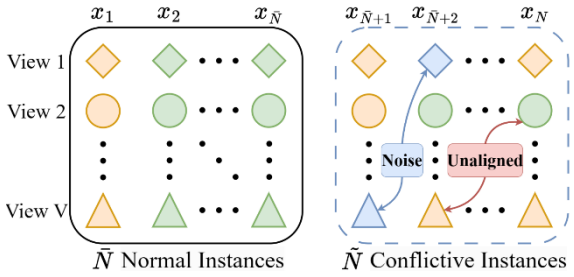


图 3　数据定义。不同的形状代表不同的视角，黄色和绿色代表两种不同的类别，蓝色代表该视角含有噪声。一个冲突的样本含了噪声数据和不对齐视角数据。

2．特定视角的证据深度学习

大多数现有的深度多视角学习方法通常依赖于在深度神经网络上使用 Softmax 层来进行分类。然而，这些基于 Softmax 的深度神经网络在准确估计预测不确定性方面存在局限性。这是因为 Softmax 只提供了预测分布的单点估计，在出现错误预测时，仍会产生高置信度的预测输出。本文通过引入证据深度学习来解决 Softmax 高置信度预测问题。

对于一个K分类问题，假设某个样本的一个特定视角$(\boldsymbol{x}_n^v, y_n)$所对应的多项式观点为$\boldsymbol{w}(\boldsymbol{b}, u, \boldsymbol{a})$，其中$\boldsymbol{b}=(b_1,\ldots,b_K)$是信念质量向量，$u$是不确定性质量，$\boldsymbol{a}=(a_1,\ldots,a_K)$是每个类别的先验概率。主观逻辑[34]要求信念质量和不确定性质量是非负的，且两者元素之和为 1，即

$$u+\sum_{k=1}^{K} b_k = 1$$

其中，$u \geq 0$，$b_k \geq 0$。概率是多项式观点的投影，如下所示：

$$p_k = b_k + a_k u.$$

通常，先验概率$\boldsymbol{a}$是根据先验知识人为手动设置的。例如，一种常见的方法是假定每个类别的先验概率相等，即：$a_k = 1/K$。

根据主观逻辑，将狄利克雷分布作为最终的类别分布的先验分布。它可以模拟二阶不确定性，而 Softmax 层中的概率值只能捕捉一阶不确定性。具体的狄利克雷分布的概率密度函数$D(\boldsymbol{p}|\boldsymbol{\alpha})$为：

$$D(\boldsymbol{p}|\boldsymbol{\alpha}) = \begin{cases} \dfrac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K} p_k^{\alpha_k - 1} & for\ \boldsymbol{p} \in \boldsymbol{S}_K \\ 0 & otherwise \end{cases}$$

其中，$\boldsymbol{p} = (p_1, \ldots, p_K)$是每个类别的概率，$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$是狄利克雷分布的参数，$\boldsymbol{S}_K$是$K$维单纯性，具体如下所示：

$$\boldsymbol{S}_K = \left\{ \boldsymbol{p} \middle| \sum_{k=1}^{K} p_k = 1,\ p_k > 0 \right\},$$

$B(\boldsymbol{\alpha})$是K维的 Beta 分布。

在本文中，证据指的是使用证据神经网络$f(\cdot)$从特定视角输入中收集的支持分类的指标$\boldsymbol{e} = f(\boldsymbol{x}_n^v)$，进而可以获得狄利克雷分布的参数$\boldsymbol{\alpha} = \boldsymbol{e} + \boldsymbol{1}$。在狄利克雷分布和多项式观点之间存在一个映射，具体如下：

$$b_k = \frac{\alpha_k - 1}{S}, u = \frac{K}{S}$$

其中，$S = \sum_{k=1}^{K} \alpha_k$。

根据上述公式可以发现，整体的证据量越少，不确定性就越高。另外，第k个类别的期望概率的计算公式如下：

$$p_k = \frac{\alpha_k}{S} = \frac{e_k + 1}{S}.$$

3．冲突观点的聚合策略

冲突的多视角数据主要分为两部分，一部分是噪声数据，一部分是未对齐数据。其中噪声数据会显示出很高的不确定性，因此需要在聚合阶段减少它们的影响；而未对齐视角将提供高度冲突的观点，并且它们的不确定性很低，这表明某些视角是不可靠的。在这种情况下，我们很难判断哪些视角的信息是可靠的。事实上，多视角学习结果的不确定性不应该随着视角数量的增加而减少，而应该与待融合的视角的质量有关，特别是当两种视角的学习结果冲突时。为了解决这一问题，本节提出了一种新的冲突观点的聚合方法。具体定义如下：

定义 1 冲突观点的聚合方法 假设$\boldsymbol{w}^A = (\boldsymbol{b}^A, u^A, \boldsymbol{a}^A)$，$\boldsymbol{w}^B = (\boldsymbol{b}^B, u^B, \boldsymbol{a}^B)$分别是同一个样本中的视角 A 和视角 B 的观点，$\boldsymbol{w}^{A\underline{\Diamond}B} = (\boldsymbol{b}^{A\underline{\Diamond}B}, u^{A\underline{\Diamond}B}, \boldsymbol{a}^{A\underline{\Diamond}B})$表示视角 A 和视角 B 观点的聚合观点，详细的计算方式如下：

$$\boldsymbol{w}^{A\underline{\Diamond}B} = \boldsymbol{w}^A \underline{\Diamond} \boldsymbol{w}^B = (\boldsymbol{b}^{A\underline{\Diamond}B}, u^{A\underline{\Diamond}B}, \boldsymbol{a}^{A\underline{\Diamond}B})$$

$$b_k^{A\underline{\Diamond}B} = \frac{b_k^A u^B + b_k^B u^A}{u^A + u^B},$$

$$u^{A\underline{\Diamond}B} = \frac{2u^A u^B}{u^A + u^B}, \boldsymbol{a}^{A\underline{\Diamond}B} = \frac{\boldsymbol{a}^A + \boldsymbol{a}^B}{2}.$$

这种聚合是通过使用多项式观点和狄利克雷分布之间的双射映射将证据映射到信念质量来实现的。从本质上讲，该聚合规则确保了聚合后的观点的质量与被聚合观点的质量成正比。具体地说，当聚合一个高度不确定的观点时，新观点的不确定性大于原来的观点；反正，则小于原观点。冲突观点的聚合方法可以简化为证据的平均，后面中将给出更详细的解释。

根据定义 1，可以用以下规则融合不同观点来获得最终的共同观点：

$$\boldsymbol{w} = \boldsymbol{w}^1 \underline{\Diamond} \boldsymbol{w}^2 \underline{\Diamond} \ldots \underline{\Diamond} \boldsymbol{w}^V.$$

根据上述融合规则，可以得到最终的多视角联合观点$\boldsymbol{w}$，从而得到最终的各类概率和总体不确定性。

4．冲突度

所提出的方法还致力于:1)在训练阶段(使用正常样本)确保模型在不同视角下的一致性;2)对视角信息冲突的程度有一个直观的感受。因此，本文引入了根据观点熵建立的冲突度(Conflict Degree)度量，具体定义如下：

定义 2 冲突度 假设$\boldsymbol{w}^A$，$\boldsymbol{w}^B$分别是同一个样本中视角 A 和视角 B 的观点，那么视角 A 和视角 B 的冲突度$c(\boldsymbol{w}^A,\boldsymbol{w}^B)$为：

$$c(\boldsymbol{w}^A,\boldsymbol{w}^B)=c_p(\boldsymbol{w}^A,\boldsymbol{w}^B)\cdot c_c(\boldsymbol{w}^A,\boldsymbol{w}^B),$$

其中$c_p(\boldsymbol{w}^A,\boldsymbol{w}^B)$是视角 A 和视角 B 之间的投影距离，$c_c(\boldsymbol{w}^A,\boldsymbol{w}^B)$是视角 A 和视角 B 之间的联合确定性。两者可以通过以下公式计算获得：

$$c_p(\boldsymbol{w}^A,\boldsymbol{w}^B)=\frac{\sum_{k=1}^{K}\left|p_k^A-p_k^B\right|}{2},$$
$$c_c(\boldsymbol{w}^A,\boldsymbol{w}^B)=(1-u^A)(1-u^B).$$

根据上述定义可以看出，该指标保证了两件事:(1)当观察到相同的预测概率分布时，$c=0$表明观点不冲突;(2)当绝对观点存在但预测概率分布不同时，$c=1$。特别地，当$c_c=0$时，它表示在一个或两个视角中存在不可靠的条件。另一方面，当$c_c=1$时，表示两者的观点都被认为是可信的，也就是说它们的不确定性为零。

5．讨论与分析

在本节中，将从理论上分析 ECML 的优势，特别是冲突多视角数据的冲突观点聚合策略。具体细节如下：

命题 1 冲突观点的聚合方法可以简化为证据的平均。

证明 1 假设$\boldsymbol{w}^A=(\boldsymbol{b}^A,u^A,\boldsymbol{a}^A)$，$\boldsymbol{w}^B=(\boldsymbol{b}^B,u^B,\boldsymbol{a}^B)$分别是同一个样本中的视角 A 和视角 B 的观点，$\boldsymbol{w}=(\boldsymbol{b},u,\boldsymbol{a})$是视角 A 和视角 B 观点的聚合后的观点。同样的e_k^A，e_k^B和e_k分别是视角 A、视角 B 和聚合后观点对应的第k个类别的证据。

$$\begin{aligned}e_k&=b_kS=\frac{b_kK}{u}\\&=\frac{b_k^Au^B+b_k^Bu^A}{u^A+u^B}\cdot\frac{u^A+u^B}{2u^Au^B}\cdot K\\&=\frac{K}{2}\cdot\frac{\frac{Ke_k^A}{S^AS^B}+\frac{Ke_k^B}{S^AS^B}}{\frac{K}{S^A}\cdot\frac{K}{S^B}}\\&=\frac{e_k^A+e_k^B}{2}\end{aligned}$$

基于这一命题，在多视角融合阶段，该方法建立了简单有效的平均池化融合层来实现冲突观点聚合。

命题 2 对于冲突观点的聚合策略，如果新观点的不确定性小于原观点的不确定性，则聚合后观点的不确定质量小于原观点;反义大于原观点。

证明 2 假设$\boldsymbol{w}^o$，$\boldsymbol{w}^a$分别是同一个样本中的两个观点，$\boldsymbol{w}^o$是原观点，$\boldsymbol{w}^a$是待聚合的新观点，$\boldsymbol{w}$是聚合后的观点。同样的u^o，u^a和u分别是对应的不确定性。

$$u=\frac{2u^ou^a}{u^o+u^a}=\frac{1}{\frac{1}{2}\left(1+\frac{u^o}{u^a}\right)}\cdot u^o.$$

由此可得聚合后观点的不确定性与待聚合观点的不确定性之间的关系：

$$\begin{cases}u<u^o, & u^a<u^o\\u=u^o, & u^a=u^o\\u>u^o, & u^a>u^o\end{cases}$$

现有的大多数可信多视角学习方法都是基于以下假设：“将另一种观点整合到原观点中，得到的不确定性质量会减小。”本文认为这是不合理的，因为:1)当整合一个可靠的视角时，融合过程应该减少整体的不确定性;2)当纳入不可靠或冲突的观点时，融合应增加不确定性。此外，现有的方法往往忽略了由于数据不一致或模型跨不同视角的性能变化引起的从不同观点收集的观点之间发生冲突的可能性。

为了更加清楚的解释这个问题，本节以一个现实中的场景为例。假设现在有两个观察者 A 和 B，他们观察从一个盒子里抽出的彩色球。这些球有四种颜色:黑、白、红、绿。假如观察者 B 是色盲，难以区分红色和绿色的球，而能够区分其他颜色的组合。另一方面，观察者 A 有正常的视力，能识别出正确的颜色。当一个红球被选中时，观察者 A 通常会认为它是红色的，而观察者 B 可能会认为它是绿色的。此时，观察者 A 和 B 对同一物体产生了不同的观点，但是两者都坚信自己的判断。假设最初不知道其中一个观察者是否是色盲，那么他们的观点都会被认为是同样可靠的。然而，现有的融合方法在结合两个观察者的观点后，会错误地降低整体的不确定性。在这个例子中，观察者可以被看作是收集多视角数据的传感器，他们的判断能力可以被视为每个视角所对应的模型的分辨能力。由于，信息收集方式的错误或者

模型本身能力的不足，最终导致了不同的视角会产生冲突的多视角信息。综上所述，在视角冲突的情况下，应该平等地对待来自每个视角的信息，考虑到观点冲突的可能性，而不能简单地降低不确定性。

6．损失函数

Softmax 函数为样本的类概率提供了一个点估计，但不提供相关的不确定性。而多项式观点或等价的狄利克雷分布可以用来模拟类概率的概率分布。因此，可以通过设计并训练神经网络，以形成对应的多项式观点。具体而言，将传统基于神经网络的分类器的 Softmax 层替换为激活函数层(如：Relu)，以确保网络输出非负值，将其作为证据向量$\boldsymbol{e}$，从而得到狄利克雷分布的参数。

首先，对于特定的单视角来说，在传统的基于神经网络的分类器中，通常采用交叉熵损失函数。因此，本文对交叉熵损失函数进行调整以适用于本文提出的方法，具体如下：

$$\begin{aligned} L_{ce}(\boldsymbol{\alpha}) &= \int \left[\sum_{j=1}^{K} -y_j \log(p_j)\right] \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_j^{\alpha_j - 1} d\boldsymbol{p} \\ &= \sum_{j=1}^{K} y_j \left(\psi\left(\sum_{j=1}^{K} \alpha_j\right) - \psi(\alpha_j)\right) \end{aligned}$$

对一批训练样本的损失可以通过对批次中每个样本的损失之和来计算。在训练过程中，模型可以发现数据中的模式，并根据这些模式为特定的类标签生成证据，以最小化总体损失。例如，当模型发现手写数字图像上存在一个大的圆形图案时，类别零的证据会增加。虽然上述损失函数可以保证每个样本的正确标签比其他类别产生更多的证据，但不能保证错误标签产生更少的证据。换句话说，如果样本不能被正确分类，那么所有证据都应倾向于零。对于均匀的狄利克雷分布，即：$S = K$，通过计算可以发现其所对应的证据量为 0，也就是说整体不确定性$u = 1$。因此，通过引入 Kullback-Leibler (KL) 散度作为正则化来惩罚无法正确分类的样本。这个正则化项为：

$$\begin{aligned} &KL[D(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}}) \parallel D(\boldsymbol{p}|\mathbf{1})] \\ &= \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \tilde{\alpha}_k\right)}{\Gamma(K) \prod_{k=1}^{K} \Gamma(\tilde{\alpha}_k)}\right) + \sum_{k=1}^{K} (\tilde{\alpha}_k - 1)\left[\psi(\tilde{\alpha}_k) - \psi(\tilde{S})\right] \end{aligned}$$

其中，$\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (\mathbf{1} - \boldsymbol{y}) \odot \boldsymbol{\alpha}$是调整后的狄利克雷分布参数，$\tilde{S} = \sum_{k=1}^{K} \tilde{\boldsymbol{\alpha}}_k$，$\Gamma(\cdot)$是伽马函数，$\lambda_t = \min(1.0, t/10) \in [0,1]$是退火系数，$t$是当前训练的次数。

因此，给定观点对应的狄利克雷分布的参数$\boldsymbol{\alpha}$，则对应的单个观点的损失函数为：

$$L_{acc}(\boldsymbol{\alpha}) = L_{ce}(\boldsymbol{\alpha}) + \lambda_t KL[D(\boldsymbol{p}|\tilde{\boldsymbol{\alpha}}) \parallel D(\boldsymbol{p}|\mathbf{1})].$$

其次，采用了最小化观点冲突程度的方法来保证在训练过程中不同观点之间结果的一致性。对于第n个样本$\{x_n^v\}_{v=1}^V$，其一致性损失的计算公式如下：

$$L_{con}^n = \frac{1}{V-1} \sum_{p=1}^{V} \left(\sum_{q \neq p}^{V} c(\boldsymbol{w}_n^p, \boldsymbol{w}_n^q)\right).$$

其中V是总视角数量，$\boldsymbol{w}_n^p$是第n个样本第p个视角的观点，$c(\boldsymbol{w}_n^p, \boldsymbol{w}_n^q)$是视角$p$和视角$q$的冲突度。

最后，最终的损失函数如下所示：

$$L = \frac{1}{N} \sum_{n=1}^{N} \left(L_{acc}(\boldsymbol{\alpha}_n) + \beta \sum_{v=1}^{V} L_{acc}(\boldsymbol{\alpha}_n^v) + \gamma L_{con}^n\right).$$

其中$\boldsymbol{\alpha}_n^v$和$\boldsymbol{\alpha}_n$分别是第n个样本的第v个视角和最终视角的狄利克雷分布的参数，β和γ是两个超参数。

四、实验结果与分析

为了验证所提出模型的有效性，本研究在六个多视角数据集上进行了实验，这些数据集的详细信息如下：

(1)HandWritten[35]包含了 2000 个从“0”到“9”的手写数字样本，每个类有 200 个样本，数据集提供了 6 种类型的特征：Pix、Fou、Fac、ZER、KAR 和 MOR。

(2)CUB[36]是一个包含 200 个类别的鸟类数据集，一共 11788 个样本。每个样本都含有一张图片和一段文本描述，使用 GoogleNet 提取图片特征和 doc2vec 提取文本特征作为两个视角。

(3) HMDB[37]是一个大规模的人类动作识别数据集，包含来自 51 个动作类别的 6718 个样本，提取 HOG 和 MBH 特征作为该数据集的两个视角。

(4) Scene15[38]包括来自 15 个室内和室外场景类别的 4485 张图像，分别提取了三种类型的特征 GIST、PHOG 和 LBP 作为不同的视角。

(5)Caltech101[39]包含了 8677 张图片，一共 101 个类别。本节选取了前十个类别，并使用 DECAF 和

VGG19 模型提取两种深度特征作为不同的视角。

（6）PIE[40]包含 680 个样本，一共 68 个类，分别提取强度、LBP 和 Gabor 特征作为 3 个视角。

数据集	样本数	类别数	视角维度
HandWritten	2000	10	240/76/216/47/64/6
CUB	11788	10	1024/300
HMDB	6718	51	1000/1000
Scene15	4485	15	20/59/40
Caltech101	8677	101	4096/4096
PIE	680	68	484/256/279

表 1　多视角数据集

本文将所提出的方法与以下两大类方法进行比较：

基于特征融合的多视角学习方法：

（1）DCCAE[41]是经典的多视角学习方法，使用归一化谱聚类计算样本的类隶属度和使用广义特征值方法计算特征融合的映射，以利用同一类样本之间的相关信息，从而寻找数据的共同表征。

（2）CPM-Nets[42]是一种多视角特征融合方法，其重点是学习一种通用的表示来处理不同视角之间的复杂关联。

（3）DUA-Nets[23]是一种不确定性感知方法，它利用反向网络将来自不同观点的内在信息整合为统一的表示。

基于决策融合的多视角学习方法：

（4）TMC[29]是多视角不确定性感知方法的先驱，它解决了不确定性估计问题，产生了可靠的分类结果。

（5）TMDL-OA[31]是一种基于证据深度学习的 SOTA 多视角决策融合方法，该方法提出了一致性度量损失来实现可靠的学习结果。

为了创建具有视角冲突样本的测试集，本文进行以下转换：（1）对于噪声视角，选取一定比例的样本，添加不同标准差的高斯噪声来制造噪声数据。（2）对于不对齐的视角，选取一定比例的样本，并随机修改某个视角的信息，使得该视角信息与其他视角描述的不一致。本章实验是基于 PyTorch 深度学习框架实现的，并对每种方法进行了 10 次运行，统计平均值和标准差。

1．对比试验

为了模型的有效性，在正常测试集和冲突测试集上对所提出的方法的性能进行验证，使用分类准确率作为模型性能的评价指标。根据以下方式创建冲突的多视角数据集：（1）对于噪声视角，选取一半的样本，添加标准差为 1 的高斯噪声来制造噪声数据。（2）对于不对齐的视角，选取一半的样本，并修改第一个视角的信息，使得该视角信息与其他视角描述的不一致。实验结果如表 2 和表 3 所示。

可以观察到，本文提出的方法在在原始数据集上的分类准确率优于所有基线方法。在 HMDB 数据集上，与第二好的(TMDL-OA)模型相比，ECML 的准确率提高了大约 2.64%;在 PIE 数据集上，与第二好的(TMDL-OA)模型相比，ECML 的准确率提高了大约 2.38%。在原始测试集上，ECML 优于所有其他对比方法，其原因可能是由于一致性损失的加入，增强了模型的学习能力，在消融研究中也证实了该猜测。

根据实验结果可以观察到，在冲突测试集上进行测试时，所有方法的准确率均明显降低。尽管如此，由于冲突观点的聚合策略，ECML 显示出对特定视角的冲突的高效的感知能力，从而在所有数据集中产生了令人印象深刻的结果。这也突出了 ECML 对于冲突多视角数据的有效性。在 PIE 数据集上，与第二好的(TMDL-OA)模型相比，ECML 的准确率提高了大约 15.84%;Scene15 数据集上，与第二好的(TMDL-OA)模型相比，ECML 的准确率提高了大约 8.55%此外，从实验结果中还发现了一些现象：（1）基于决策融合的方法优于基于特征融合的方法。这可能是由于面对视角冲突，当使用基于特征融合的方法时，融合后的特征可能会消失，导致最终决策失败。当使用基于决策融合的方法时，每个单独视角的决策都是正常的。尽管最终的决定可能是错误的，但它们比基于特征融合的方法要好。（2）不同数据集的性能差异很大。原因可能是有些数据集(PIE, Scene15)在不同类别之间的特征差异很大，而有些数据集(CUB,

数据集	DCCAE	CPM-Nets	DUA-Nets	TMC	TMDL-OA	**Ours**
HandWritten	95.45 ± 0.35	94.55 ± 1.36	98.10 ± 0.32	98.51 ± 0.13	99.25 ± 0.45	**99.40 ± 0.00**
CUB	85.39 ± 1.36	89.32 ± 0.38	80.13 ± 1.67	90.57 ± 2.96	95.43 ± 0.20	**98.50 ± 2.75**
HMDB	49.12 ± 1.07	63.32 ± 0.43	62.73 ± 0.23	65.17 ± 2.42	88.20 ± 0.58	**90.84 ± 1.86**
Scene15	55.03 ± 0.34	67.29 ± 1.01	68.23 ± 0.11	67.71 ± 0.30	75.57 ± 0.02	**76.19 ± 0.12**
Caltech101	89.56 ± 0.41	90.35 ± 2.12	93.43 ± 0.34	92.80 ± 0.50	94.63 ± 0.04	**95.36 ± 0.38**
PIE	81.96 ± 1.04	88.53 ± 1.23	90.56 ± 0.47	91.85 ± 0.23	92.33 ± 0.36	**94.71 ± 0.02**

表 2　模型在正常多视角测试集上的分类准确率（%）

数据集	DCCAE	CPM-Nets	DUA-Nets	TMC	TMDL-OA	**Ours**
HandWritten	82.85 ± 0.38	83.34 ± 1.07	87.16 ± 0.34	92.76 ± 0.15	93.05 ± 0.05	**94.40 ± 0.05**
CUB	63.57 ± 1.28	68.82 ± 0.17	60.53 ± 1.17	73.37 ± 2.16	74.43 ± 0.26	**76.50 ± 1.15**
HMDB	29.62 ± 1.79	42.62 ± 1.43	43.53 ± 0.28	47.17 ± 0.15	67.62 ± 0.28	**70.84 ± 1.19**
Scene15	25.97 ± 2.86	29.63 ± 1.12	26.18 ± 1.31	42.27 ± 1.61	48.42 ± 1.02	**56.97 ± 0.52**
Caltech101	60.90 ± 2.32	66.54 ± 2.89	75.19 ± 2.34	90.16 ± 2.50	90.63 ± 2.05	**92.36 ± 1.48**
PIE	26.89 ± 1.10	53.19 ± 1.17	56.45 ± 1.75	61.65 ± 1.03	68.16 ± 0.34	**84.00 ± 0.14**

表 3　模型在冲突多视角测试集上的分类准确率（%）

Caltech101)的差异相对较小。特征差异较大的数据集经过冲突处理后，冲突程度较高。ECML 可以解决这种情况，而传统的方法不能。另外，性能也可能与视角的数量有关。

2．冲突度量可视化分析

为了更加直观地感受 ECML 方法中提出的冲突度量，本节基于 HandWritten 数据集对冲突度度量进行可视化操作。为了构建冲突样本，本节修改了数据集中第一个视角中的内容，使得该视角与其他视角中的内容不对齐。下图显示了 HandWritten 数据集上六个视角的冲突程度。颜色越深代表冲突的程度越高。

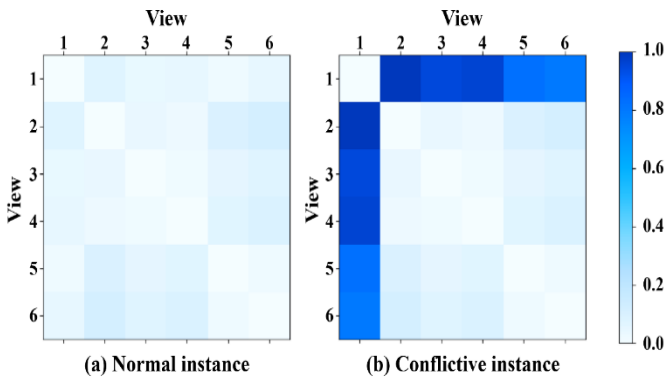


图 4　冲突度可视化分析

图 4 清楚地表明了 ECML 可以有效地捕获和量化视角之间的冲突程度，这一发现进一步验证了 ECML 的可靠性。

3．不确定性估计分析

为了进一步评估 ECML 的不确定性量化能力，本节可视化了基于 CUB 数据集的正常和冲突测试集的分布。为了构建冲突测试集，引入了标准偏差 $\sigma = 0.1$、1、5、10的高斯噪声随机加入正常的测试集中。实验结果如图 5 所示。

结果表明，当噪声强度较低($\sigma = 0.1$)时，冲突数据集的不确定性分布曲线与正常数据集的分布曲线基本一致。然而，随着噪声强度的增加，冲突数据集的不确定性也随之增加。这一发现表明，不确定性与数据集的质量相关，并且验证了本章提出的方法在不确定性量化方面的有效性。

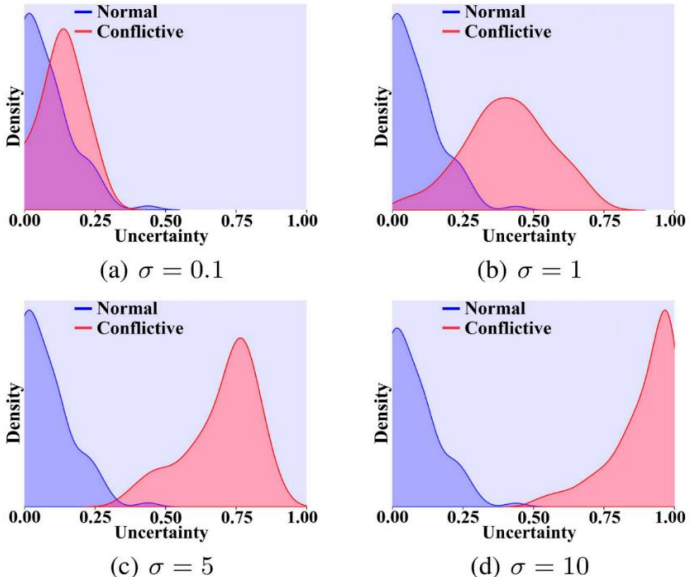


图 5　不确定性量化分析

4．消融实验

为了验证模型中一致性损失的有效性，本节在 Scene15 和 PIE 数据集上进行了消融实验，并与以下变体进行比较:

（1）ECML-RC：该变体在训练过程中去除了一致性损失

（2）ECML-V1：该变体只使用视角一的信息

（3）ECML-V2：该变体只使用视角二的信息

（4）ECML-V3：该变体只使用视角三的信息

实验根据以下方式创建冲突的多视角数据集：(1) 对于噪声视角，选取一半的样本，添加标准差为 1 的高斯噪声来制造噪声数据。(2) 对于不对齐的视角，选取一半的样本，并修改第一个视角的信息，使得该视角信息与其他视角描述的不一致。实验结果如图 6 所示。

从实验结果可以看出：(1) 采用冲突观点的聚合策略对所有视角信息进行整合相比于单独的单视角信息，其所训练出的模型的性能更好。(2) 使用一致性损失函数能够增强模型的学习能力。综上所述，消融实验的结果进一步验证了本章提出的模型中各个组成部分的有效性。冲突观点的聚合策略和一致性损失函数在多视角学习任务中各自发挥了重要作用，而将它们集成在一起则能够更全面地捕捉多视角间的冲突信息，提高多视角任务的可靠性和安全性。

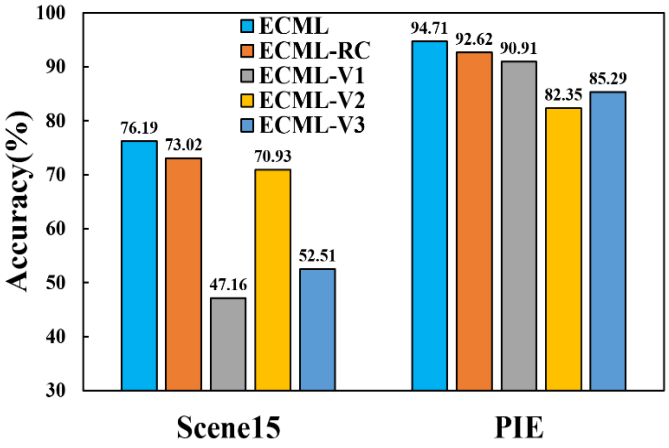


图 6　一致性损失有效性分析

四、总结

针对可靠的多视角学习问题，本文提出了一种可信冲突多视角学习算法。该方法试图形成由信念质量向量和决策可靠性组成的特定视角的观点。它通过一个简单有效的平均池化层进一步聚合冲突观点。从理论上证明了该方法可以准确地模拟多视角公共可靠度和视角特定可靠度之间的关系。此外，本文对方法进行了扩展，通过最小化观点之间的冲突程度来保证不同观点之间结果的一致性。大量实验验证了所提出的方法在性能上的优越性以及其良好的稳定性和可靠性。

责任编委　王金甲

参考文献

[1] Cai Xu, Ziyu Guan, Wei Zhao, Hongchang Wu, Yunfei Niu and Beilei Ling. Adversarial incomplete multi-view clustering. In Proceedings of International Joint Conference on Artificial Intelligence(IJCAI), 2019.

[2] Shudong Huang, Hongjie Wu, Yazhou Ren, Ivor Tsang, Zenglin Xu, Wentao Feng and Jiancheng Lv. Multi-view subspace clustering on topological manifold. In Neural Information Processing Systems(NeurIPS), 2022.

[3] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang and Jinxing Li. A survey on incomplete multiview clustering. In IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022.

[4] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. ArXiv preprint arXiv:1701.08251, 2017.

[5] Ge Fan, Chaoyun Zhang, Kai Wang, and Junyang Chen. MV-HAN: A hybrid attentive networks based multi-view learning model for large-scale contents recommendation. In Proceedings of IEEE/ACM International Conference on Automated Software Engineering, 2022.

[6] Yanchao Tan, Chengjun Kong, Leisheng Yu, Pan Li, Chaochao Chen, Xiaolin Zheng, Vicki S. Hertzberg, and Carl Yang. 4sdrug: Symptom-based set-to-set small and safe drug recommendation. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022.

[7] Paul Pu Liang, Amir Zadeh and Louis-Philippe Morency. Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. ArXiv preprint arXiv:2209.03430, 2022.

[8] Shuping Zhao, Jie Wen, Lunke Fei and Bob Zhang. Tensorized incomplete multi-view clustering with intrinsic graph completion. In Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[9] Chaoyang Zhang, Zhengzheng Lou, Qinglei Zhou and Shizhe Hu. Multi-View Clustering via Triplex Information Maximization. In IEEE Transactions on Image Processing, 2023.

[10] Jing Gao, Wei Fan, Deepak Turaga, Srinivasan Parthasarathy and Jiawei Han. A spectral framework for detecting inconsistency across multi-source object relationships. In IEEE International Conference on Data Mining, 2011.

[11] Alejandro Marcos Alvarez, Makoto Yamada, Akisato Kimura, and Tomoharu Iwata. Clustering-based anomaly detection in multi-view data. In Proceedings of ACM international conference on Information & Knowledge Management, 2013.

[12] Sheng Li, Ming Shao and Yun Fu. Multi-view low-rank analysis for outlier detection. In Proceedings of SIAM International Conference on Data Mining, 2015.

[14] Handong Zhao, Hongfu Liu, Zhengming Ding and Yun Fu. Consensus regularized multi-view outlier detection. In IEEE Transactions on Image Processing, 2017. Kai Li, Sheng Li, Zhengming Ding, Weidong Zhang and Yun Fu. Latent discriminant subspace representations for multi-view outlier detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[15] Dongdong Hou, Yang Cong, Gan Sun, Jiahua Dong and Jun Li; Kai Li. Fast multi-view outlier detection via deep encoder. In IEEE Transactions on Big Data, 2020.

[16] Lampert C H, Krömer O. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In ECCV 2010.

[17] Xianchao Zhang, Linlin Zong, Xinyue Liu and Hong Yu. Constrained NMF-based multi-view clustering on unmapped data. In Proceedings of the AAAI Conference on Artificial Intelligence, 2015.

[18] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv and Xi Peng. Partially view-aligned clustering. In Neural Information Processing Systems(NeurIPS), 2020.

[19] Mouxing Yang, Yunfan Li, Zhenyu Huang, Zitao Liu, Peng Hu and Xi Peng. Partially view-aligned representation learning with noise-robust contrastive loss. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition(CVPR), 2021.

[20] Xianchao Zhang, Mengyan Chen, Jie Mu and Linlin Zong. Adaptive View-Aligned and Feature Augmentation Network for Partially View-Aligned Clustering. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2023.

[21] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. In Artificial Intelligence Review, 2023.

[22] Murat Sensoy, Lance Kaplan and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Neural Information Processing Systems(NeurIPS), 2018.

[23] Yu Geng, Zongbo Han, Changqing Zhang and Qinghua Hu. Uncertainty-Aware Multi-View Representation Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

[24] Denker John and LeCun Yann. Transforming neural-net output levels to probability distributions. In Neural Information Processing Systems(NIPS), 1990.

[25] Radford M. Neal. Bayesian training of backpropagation networks by the hybrid Monte Carlo method. Technical Report CRG-TR-92-1, Dept. of Computer Science, University of Toronto, 1992.

[26] Hinton G E, Van Camp D. Keeping the neural networks simple by minimizing the description length of the weights. In Proceedings of annual conference on Computational learning theory, 1993.

[27] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Neural Information Processing Systems(NIPS), 2017

[28] Shanmugam D, Blalock D, Balakrishnan G, et al. Better aggregation in test-time augmentation. In Proceedings of the IEEE/CVF international conference on computer vision, 2021.

[29] Zongbo Han, Changqing Zhang, Huazhu Fu and Joey Tianyi Zhou. Trusted Multi-View Classification. In International Conference on Learning Representations, 2021.

[30] Myong Chol Jung, He Zhao, Joanna Dipnall, Belinda Gabbe and Lan Du. Uncertainty estimation for multi-view data: The power of seeing the whole picture. In Neural Information Processing Systems(NeurIPS), 2022.

[31] Wei Liu, Xiaodong Yue, Yufei Chen and Thierry Denoeux. Trusted Multi-View Deep Learning with Opinion Aggregation. In Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

[32] Wei Liu, Yufei Chen, Xiaodong Yue, Changqing Zhang and Shaorong Xie. Safe Multi-View Deep Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[33] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou and Xi Peng. Provable Dynamic Fusion for Low-Quality Multimodal Data. In Proceedings of International Conference on Machine Learning, PMLR, 2023.

[34] Audun Jsang. Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated, 2018.

[35] Perkins S, Theiler J. Online feature selection using grafting. In Proceedings of International Conference on Machine Learning(ICML), 2003.

[36] He X, Peng Y. Fine-grained visual-textual representation learning. In IEEE Transactions on Circuits and Systems for Video Technology, 2019.

[37] Wishart D S, Tzur D, Knox C, et al. HMDB: the human metabolome database. In Nucleic acids research, 2007.

[38] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene categories. In IEEE computer society conference on computer vision and pattern recognition, 2005.

[39] Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In IEEE conference on computer vision and pattern recognition workshop, 2004.

[40] Tulyakov S, Fitzgibbon A, Nowozin S. Hybrid vae: Improving deep generative models using partial observations. ArXiv preprint arXiv:1711.11566, 2017.

[41] Sheng Wang, Jianfeng Lu, Xingjian Gu, Benjamin A. Weyori and Jing-yu Yang. Unsupervised discriminant canonical correlation analysis based on spectral clustering. In Neurocomputing, 2016.

[42] Changqing Zhang, Zongbo Han, yajie cui, Huazhu Fu, Joey Tianyi Zhou and Qinghua Hu. CPM-Nets: Cross partial multi-view networks. In Neural Information Processing Systems(NIPS), 2019.

徐 偲

西安电子科技大学副教授，硕士生导师。主要研究方向为可信多模态深度学习。在 IEEE TPAMI、IEEE TNNLS、IEEE TCyb、IEEE TII、NeurIPS、SIGKDD、AAAI、IJCAI、ACM MM 等中科院一区 IEEE 汇刊或 CCF A 类学术会议发表论文 22 篇，其中一作 9 篇，通讯 3 篇。主持国家自然科学基金青年基金项目等 5 项科研项目，作为核心骨干成员参与国家自然科学基金重点项目 2 项。担任 SCI 期刊 Array 编委，SCI 期刊 Sensors 和 Mathematics 的客座编委，IEEE TPAMI、IEEE TKDE、ICML、AAAI 等 20 余个中科院一区期刊或 CCF A 类会议的审稿人。

Email: cxu@xidian.edu.cn

司佳俊

西安电子科技大学计算机科学与技术学院 2021 级硕士研究生，导师为赵伟教授，主要研究方向为可信多模态深度学习。

Email: jiajunsi@stu.xidian.edu.cn

管子玉

西安电子科技大学教授，博士生导师，曾获国家级青年人才项目。主持基金委重点项目、面上项目等国家级项目。在数据挖掘、信息检索、数据管理等领域的顶级国际会议和期刊发表论文70余篇，包括TKDE、TPAMI、TNNLS、TIP、VLDB、SIGMOD、SIGIR、ICDE、WWW、AAAI、IJCAI、SIGKDD、CVPR等。担任TKDE、Neurocomputing和International Journal of Machine Learning and Cybernetics编委，担任多个高水平国际会议程序委员会委员/资深委员，如SIGKDD、IJCAI、AAAI、NeurIPS、SIGIR、CIKM等，担任领域内知名国际会议环太平洋多媒体会议PCM 2016组织主席。

Email: zyguan@xidian.edu.cn

赵 伟

西安电子科技大学教授，博士生导师。主要从事智能媒体计算和机器学习等方面的研究工作。近五年，以第一作者/通讯作者在相关领域国际著名期刊及学术会议发表论文20余篇，如：IEEE TPAMI、IEEE TKDE、NeurIPS、AAAI、SIGKDD等。担任CCF A类期刊IEEE TKDE和中科院一区期刊IEEE TNNLS编委，常年担任领域内顶级会议和期刊审稿人。研究成果获浙江省自然科学一等奖、陕西省自然科学优秀学术论文二等奖等。

Email: ywzhao@mail.xidian.edu.cn

顶会观察

AAAI 2024

中山大学 王岩 任传贤

人工智能促进协会（Association for the Advancement of Artificial Intelligence，AAAI），其每年举办的现场会议是人工智能领域里历史最悠久、涵盖内容最广泛的国际顶级学术会议之一，今年已是第 38 届。据中国计算机学会推荐国际学术会议和期刊目录，AAAI 为 CCF-A 类会议。今年大会的主席成员包括：来自牛津大学的 Michael wooldridge、来自马萨诸塞州东北大学的 Jennider Dy 和来自德克萨斯州大学的 Sriraam Natarajan。值得关注的是，AAAI 2024 大会三篇论文入选最佳论文奖，分别为：Reliable Conflictive Multi-view Learning，作者徐偲（Cai Xu）、赵伟（Wei Zhao）等来自西安电子科技大学；GxVAEs: Two Joint VAEs Generate Hit Molecules from Gene Expression Profiles，作者 Chen Li 和 Yoshihiro Yamanishi 来自名古屋大学；Proportional Aggregation of Preferences for Sequential Decision Making，作者 Nikhil Chandak、Shashwat Goel 和 Dominik Peters 分别来自海得拉巴国际信息技术学院以及巴黎第九大学。AAAI 2024 大会于 2024 年 2 月 20 日至 2024 年 2 月 27 日在加拿大温哥华举办，包括 4 天的正会和 4 天的 Workshops & Tutorial and Lab Forum。

一、 会议概况

AAAI 2024 继续在线下举办，无法线下参会人员需提前向组委会申明，也可以选择线上参会。据主办方统计，截至大会开幕，约 5100 人注册参会，其中 Technical program 注册人数约 2700 人，Technical plus workshop, bridges, labs, tutorials 注册人数约 800 人。现场超 4000 人参会，亲临现场的作者分不同场次做 presentation 包括 oral 和 poster。线上参会者只能观看，不做在线工作汇报。

大会的主席团成员（General Chairs, GC）介绍了会议的具体安排：39 场 workshops、22 场 tutorials、16 场 senior member presentation、7 场 Invited talk、3 场 panels。大会的程序主席（Program Chairs, PC）对 AAAI 2024 论文的审稿情况作了详细介绍：在审稿过程中，有 136 篇投稿违反了双重提交的原则被撤稿。在排除不同阶段被撤回的论文后，大会最终收稿数量为 10504 份。所有这些投稿平均收到了 4 条评审意见，评审意见总数超过 21600 条。大会为了确保评审尽可能公平公正，每篇论文随机分配到不同大洲的审稿人，并且各个审稿人之间相互独立。论文作者在第一阶段都有机会提交 rebuttal，然后由 AC 分配给每篇论文的审稿人进行讨论。每篇论文的最终录用决定是由 AC 与一位 SPC 讨论后作出的。此外，程序委员会主席监督整个审稿过程，尤其是在 AC 的决定与所有审稿人的意见存在明显分歧时，最终由程序委员会中多个 APC 和 PC 共同讨论并决定。除此之外，大会也建立了与 NeurIPS 的快速通道，一些投稿 NeurIPS 2024 较高分数的论文被允许直接进入 AAAI 2024 第二阶段。

二、 录用情况

AAAI 2024 大会共收到 12100 份投稿，总共有 9862 份有效投稿（main track），来自中国的学者投稿量位居榜首,占比约为 48.2%。最终有 2527 篇(24.1%)论文被接收，包括 main track 2340 篇（24.4%）。其中仅有 197 篇论文入选 oral，oral 率约为 8.4%。相较于上一届,今年 AAAI 的投稿量提升 12.4%(1085 篇)；录用论文数量提升 46.8%（806 篇）。从大会公布的数据来看，今年大会收到的有效投稿和录用数量相比上一届都有大幅提高，继续打破历史纪录，可见被接受论文的质量之高。AAAI 2024 会议涵盖的主题包括：认知建模和认知系统、计算机视觉、约束满足和优化、数据挖掘和知识管理、博弈论和经济范式、智能机器人、知识表示与推理、机器学习、多智能体系统、自然语言处理、调度与规划、不确定性推理、搜索和优化等。其中，位于前三位的领域依次是：计算机视觉、机器学习、自然语言处理。3D 计算机视觉，强化学习和大语言模型主题等是 AAAI 2024 较火热的几个关键词。国内重多科研机构和企业在 AAAI 2024 上斩获颇丰包括：腾讯、阿里、小红书、商汤等。其中，腾讯优图实验室共有 27 篇论文入选，内容涵盖表格结构识别、异常图像生成、医学图像分割等多个研究方向。中国人民大学高瓴人工智能学院有 18 篇论文被录用，内容包括信息检索、大型语言模型量化、3D 分子表征模型等多个方向。香港中文大学（深圳）理工学院有 9 篇论文入选，内容涵盖联邦学习公平性、分布式训练优化、三维密集分割等方向。国外机构比如谷歌、微软、Adobe 等也有不错的表现。

三、 热点论文

AAAI 2024 公布了最佳论文奖，共 3 篇论文入选，作者中有不少华人的身影。论文具体介绍如下：

最佳论文 1：Reliable Conflictive Multi-view Learning[1]，作者来自西安电子科技大学的团队。自 1984 年该学术会议设立最佳论文奖以来，以国内单位（含港澳台地区）为第一单位获得该奖项的第三篇论文。多视图学习的目标是通过结合多种特征，实现更全面的数据描述。现有大多数工作都假设多个视图是严格对齐的。然而，现实世界中的多视图数据可能包含低质量的冲突实例，这些实例在不同视图中都会显示冲突信息。针对此问题，以前方法主要集中在消除冲突数据实例或者替换冲突视图。然而，现实应用通常需要为冲突实例做出决策，而不仅仅是消除它们。为了解决这个问题，作者提出了一个新的可靠冲突多视图学习（RCML）问题，该问题要求模型为冲突多视图数据提供决策结果和附加的可靠性。作者开发了一种证据冲突多视图学习（ECML）方法。ECML 首先学习视图特定的证据，这可以称为从数据中收集的每个类别的支持量。然后可以构建包含决策结果和可靠性的视图特定意见。在多视图融合阶段，作者提出了一种冲突意见聚合策略，并从理论上证明了该策略可以准确地模拟多视图通用和视图特定可靠性的关系。这项工作的意义在于它有可能极大提升 AI 系统在关键应用如自动驾驶中的准确性和可信度。在这些应用中，AI 系统需要协调来自不同传感器的相互冲突的数据，从而做出安全的决策。多视图学习一直是 AI 领域的一个复杂挑战，而这篇论文提出新的可靠冲突多视图学习方法，不仅为解决这个问题迈出了重要的一步，也为自动驾驶等关键领域的未来发展提供了有力支持。

最佳论文 2：GxVAEs: Two Joint VAEs Generate Hit Molecules from Gene Expression Profiles[2]，作者来自名古屋大学的团队。作者认为从头生成具有生物活性和药物特性的类似候选药物分子是计算机辅助药物发现中的一项重要任务。现有人工智能技术可以生成具有所需化学性质的分子，但大多数研究忽略了与疾病相关的细胞环境的影响。本文提出了一种名为 GxVAEs 的新型深度生成模型，该模型利用两个联合变分自编码器（VAEs）从基因表达谱中生成类似候选药物分子。第一个 VAE，称为 ProfileVAE，从基因表达谱中提取潜在特征。提取的特征作为条件指导第二个 VAE，称为 MolVAE，生成类似候选药物分子。GxVAEs 在分子生成和生物系统中的细胞环境之间架起了一座桥梁，并产生了在特定疾病背景下具有

生物意义的分子。在生成治疗性分子的实验和案例研究中，GxVAEs 的表现优于目前最先进的基线方法，并生成了具有潜在生物活性和药物特性的类似候选药物分子。这项研究可能会极大地提高新药的研发效率。通过展现其生成候选药物分子的能力，该模型有望为目前尚无有效治疗手段的疾病带来治疗上的突破。

最佳论文 3：Proportional Aggregation of Preferences for Sequential Decision Making[3]，作者来自海得拉巴国际信息技术学院和巴黎第九大学的团队。论文解决了顺序决策中公平性的复杂问题，具体研究了给定选民偏好的公平序贯决策问题。在每一轮中，决策规则必须从一组备选方案中选择一个决策，其中每位选民报告备选方案中的选择。目标不是选择每一轮中最受欢迎的选择，而是寻找比例代表。作者使用基于比例正当代表（PJR）的公理来正式化这一目标，这些公理是在多胜者投票相关文献中提出的，最近被应用于多问题决策。作者证明了三个有吸引力的投票规则满足这种风格的公理。其中一个可以在线做出决策，另外两个满足公理的加强版本，但可以做出半在线或完全离线决策。前两个规则可以在多项式时间内计算，而第三个规则则涉及基于 NP 难的优化问题，但它允许满足相同公理性质的多项式时间局部搜索算法。作者还基于合成数据和美国政治选举的结果，给出了这些规则性能的经验结果。这项研究将有助于人们日益认识到 AI 开发中道德考虑的重要性。

四、 大会获奖

AAAI 经典论文奖：旨在表彰被认为最具影响力的论文，这些论文是从特定会议年份中挑选出来的。2024 年的奖项将颁发给第二十三届人工智能会议中最具影响力的论文。《Maximum Entropy Inverse Reinforcement Learning》[4]，作者来自卡内基梅隆大学，这项 2008 年的研究将熵正则化引入强化学习，从而提高了预测、模仿学习、决策和人类-AI 对齐的预测准确率。研究表明将模仿学习问题设计为马尔可夫决策问题（Markov Decision Problems）的解决方案是有益的。这种方法将学习过程简化为恢复效用函数的问题。这项工作开发了一种基于最大熵原理的概率方法，其在决策序列上提供了一个定义良好的全局规范化分布，同时提供了与现有方法相同的性能保证。研究者在对现实世界的导航和驾驶行为建模的背景下开发技术，收集的数据本质上是嘈杂且不完美的。本文提出的概率方法可以对路线偏好进行建模，并提供了一种基于部分轨迹推断目的地和路线的强大新方法。

AAAI 杰出服务奖：旨在每年表彰对人工智能社区做出非凡贡献的个人。服务的领域可能包括但不限于社会服务，担任编辑，组织会议，在其他组织（如计算机研究协会、美国计算机学会或电气电子工程师协会）中代表人工智能，或作为政府机构合同监督员或项目主任提供具有影响力的服务，对人工智能领域产生积极影响。被提名者必须是 AAAI 的当前成员。今年的获奖者是 Ashok Goel，其在人工智能领域的杰出服务，特别是在担任《AI 杂志》主编和《交互式 AI 杂志》创始编辑的过程中，持续在人工智能教育的跨学科研究方面做出的贡献而荣获此奖。

AAAI 人工智能促进人类福祉奖：旨在表彰人工智能在保护、增强和改善人类生活方面产生长期积极影响。今年获得这一奖项的是哈佛大学计算机科学教授 Milind Tambe，大会表彰其将新型人工智能技术开创性地应用于公共安全、公共保护、公共卫生领域，对国际范围内的人类产生重大益处的贡献。Milind Tambe 是 AAAI Fellow、ACM Fellow，他也担任 Google Research「AI for Social Good」计划的负责人。

AAAI/EAAI Patrick Henry Winston 杰出教育家奖：颁发给在人工智能教育领域做出重大贡献、为人工智能社区带来长期益处的个人（或团体）。该奖项以具有开创性和启发性的 AI 教育家 Patrick Henry Winston 的名字命名，来纪念他的贡献。今年的获奖者是 Charles Isbell（来自威斯康星大学麦迪逊分校）和 Michael L. Littman（来自布朗大学），大会表彰他们通过在线课程为成千上万的学生提供人工智能和机器学习的创新教学，并通过富有创意

和娱乐性的方式向公众普及相关知识，因而荣获此奖。

五、 总结展望

AAAI 2024 大会中大语言模型应用、扩散模型分析、标签学习、联邦学习、强化学习、提示学习等领域保持较高热度。综合近年来 AAAI 收录论文情况来看，会议越来越重视实际问题的应用，针对数据孤岛以及隐私保护问题，研究联邦学习；针对序贯决策问题，研究强化学习；解决数据分类和标注问题，研究标签学习，部分标签学习以及提示学习；提示学习可以有效地提升预训练的大型语言模型解决各种自然语言处理任务性能。这种方法通过向模型提供有效的“提示”（prompt），使模型能够理解并解决新的任务，而无需进行繁琐的微调（finetuning）过程，逐步提升深度模型泛化性及环境适应性。笔者认为人工智能技术发展的目的是能解决工业应用和人们生活中的实际需求，因此如何推动模型落地应用，如何让模型从识别到理解、从大数据驱动到小样本、零样本学习、从单一静态到多模态多应用场景，是人工智能未来的发展方向，也是开展科研工作的重要落脚点。

责任编委 崔海楠

参考文献

[1] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, Xiyue Gao. Reliable Conflictive Multi-view Learning. AAAI 2024.

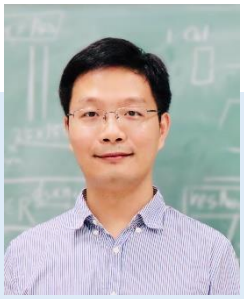
[2] Chen Li, Yoshihiro Yamanishi. GxVAEs: Two Joint VAEs Generate Hit Molecules from Gene Expression Profiles. AAAI 2024.

[3] Nikhil Chandak, Shashwat Goel, Dominik Peters. Proportional Aggregation of Preferences for Sequential Decision Making. AAAI 2024.

[4] Brian D. Ziebart, Andrew Maas, J.Andrew Bagnell, and Anind K. Dey. Maximum Entropy Inverse Reinforcement Learning. AAAI 2008.

王 岩

中山大学数学学院 2021 级硕士研究生，主要研究方向为计算机视觉与机器学习。
Email：wangy2277@mail2.sysu.edu.cn



任传贤

中山大学数学学院教授、副院长，科学计算与计算机应用系主任，CCF-CV 专委会副秘书长，中国数学会计算数学分会常务理事，广东省（广州）工业与应用数学学会副理事长兼秘书长。长期关注高维异构视觉数据的表征迁移学习算法，在国际重要学术期刊 IEEE TPAMI、TIP、TNNLS、TMI、MedIA 以及重要学术会议 CVPR、AAAI 等发表论文五十余篇，曾获得 2015 年度教育部自然科学二等奖和 2022 年度中国图象图形学学会自然科学二等奖。
Email: rchuanx@mail.sysu.edu.cn

厦门大学严严教授访谈

2024 年 2 月 29 日，《CCF-CV 专委简报》在线采访了厦门大学博士生导师严严教授。下面是采访实录。

问题 1：严老师，您好！首先，请您分享一下您的个人学习和研究经历。

谢谢您的采访。我本科在电子科技大学电子工程学院学习信息对抗技术。本科毕业后保送到清华大学电子工程系图像所读博，从事计算机视觉方面研究。在读博士期间，我有幸在章毓晋教授的指导下深入研究了人脸分析算法。博士毕业后，为了进一步提高自己的工程能力，我选择在日本东京的诺基亚研发中心以及新加坡的松下研究院工作，先后带领团队开发了大规模人脸聚类算法和家庭影院视频分析算法，积累了不少的工业界经验。这些经历不仅加深了我的理论知识，也使我在实践中得以运用和完善所学。2011 年后，考虑到自己的职业发展和国内的良好环境，我辞去企业的工作选择回国任教。目前，我在厦门大学信息学院计算机系从事教学和科研工作，致力于将我的学术和行业经验传授给学生，并继续推动计算机视觉领域的前沿研究。

问题 2：您曾在日本东京 NOKIA（诺基亚）研发中心（NOKIA 全球六大研发中心之一）图像软件组（Imaging Software Team）工作，担任算法助理研究员，也曾在新加坡 PANASONIC（松下）研究院（PANASONIC 海外最大研究院）多媒体处理组（Media Processing Group）工作，担任项目负责人，能分享一下您在这些国外核心研究机构的研究经历及您的感悟么？与国内的一些研究机构相比，你觉得他们的不同之处，或者值得借鉴之处有哪些？

我在日本诺基亚研发中心图像软件组从事人脸聚类算法的研究。该图像软件组是一个跨国的、高度专业化的团队，聚集了来自世界各地的优秀人才（包括来自中国、新加坡、英国、美国、韩国、芬兰和日本等国家的同事），他们在嵌入式图像算法开发方面具有深厚的专业知识和经验（应用到当时诺基亚各类手机中）。在这里，我有机会接触到先进的图像处理技术和工具，参与了项目的开发和优化，极大拓展了我的视野和技能，特别是我们开发的一款以人脸聚类为核心的图像浏览器曾经获得诺基亚最佳用户推荐奖。与国内研究机构相比，诺基亚注重团队合作和创新，非常鼓励员工提出新想法并积极探索解决方案。在松下研究院的多媒体处理组，我有机会领导项目团队，负责整个项目的规划、实施和交付，这提升了我的项目管理能力。松下注重技术创新和市场导向，重视产品的实用性和用户体验，这一点在项目开发中表现得尤为突出。我感觉与国内研究机构相比，国外研究机构注重自由和开放的工作氛围，鼓励员工自主探索和创新，这为团队的发展提供了良好的环境。当然，国内的科研机构在团队管理、项目实施和技术创新方面也有着独特的优势。

问题 3：您曾多次担任国际会议本地主席、程序委员会主席，还担任了很多权威期刊的审稿专家，请问您从这些经历中获得的经验和感悟是什么？

担任国际会议本地主席、程序委员会主席以及权威期刊的审稿专家是我学术生涯中的重要组成部分，这些经历让我受益匪浅。我可以深入了解学术界的运作机制。

通过担任会议本地主席，我深入了解了学术会议的组织和管理流程，包括会议议程的制定、论文提交与审稿、程序安排、专家接待、住宿安排等各个环节。这让我对学术界的运作机制有了更清晰的认识。在组织大型学术会议的过程中，我需要与各方沟通协调，领导团队完成各项工作。这锻炼了我的团队协作和管理能力。而作为审稿专家，我们往往需要对提交的论文进行严格评审，确保其质量和学术水平。这让我更加关注学术研究的原创性、创新性和科学性，加深了对学术质量的认识和要求。而且可以了解同行的研究进展，拓展了自己的学术视野。

问题 4：您获得过福建省科学技术进步奖一等奖、厦门市科学技术进步奖三等奖和清华大学电子工程系学术新秀奖等多项奖项，请问能否介绍一下您的这些获奖经历，分享一下您是如何做出这些成就的？

这个是团队一起努力的结果。福建省科学技术进步奖一等奖、厦门市科学技术进步奖的获奖项目都是计算机视觉在智慧城市和智慧交通中的一些具体的应用，比如我们把目标检测、人脸分析、行为和轨迹分析算法通过和公司的产学研合作，进行成果转化和产品化，得到了市场的高度认可。这些奖项的背后是对我们整个团队多年来科研工作的不懈努力、创新实践的认可。通过与其他优秀同事的合作，可以不断积累经验、开展研究、推动项目进展，最终取得了一系列突出的研究成果，为科技进步和社会发展做出贡献。

问题 5：您的众多研究成果中，请问哪一项是您认为最值得骄傲的？

在众多研究成果中，让我最为满意的应该还是人脸分析方面的一系列工作，我们从小样本学习、增量学习、联邦学习、解耦表示等不同的角度去研究开放环境下的人脸分析性能表现，一系列工作发表在 IEEE TPAMI、IJCV、IEEE TIP、TIFS、TAFFC、CVPR、ECCV、ACM MM 等期刊和会议上。我觉得在一个方向上持续探索是有价值的。当然，我觉得作为一名老师最开心的就是自己的学生能够取得优异的科研成绩（包括多名学生获得福建省优秀研究生论文），所以我感觉最令我骄傲的就是，自己的学生能够在各行各业为国家的人工智能战略做出自己的积极贡献。

问题 6：您在 TPAMI、IJCV 等著名期刊和 CVPR、ICCV 等国际著名会议上发表了论文 100 余篇，请问您是如何做到如此高产出的？能给其他老师和研究生们分享一些发表高水平论文的经验和建议么？

惭愧！和国内顶级研究机构相比，我们还是存在不少差距。我们首先选择一个有足够研究空间和潜力的研究方向，这是发表高水平论文的基础。要密切关注领域前沿，找准研究热点和问题，选择一个既能够挑战自己又有实际应用价值的课题。在开始研究之前，要对相关领域的文献进行深入阅读和调研，了解已有研究成果和方法，找到研究的切入点和价值所在，避免重复造轮子。精心撰写和修改论文。撰写论文是研究工作的重要环节，要注意逻辑严谨、语言通顺，确保论文结构清晰、内容完整。在修改论文时，要充分吸收审稿意见和建议，不断完善和提高论文质量，保持持续学习和积累经验。科研是一个不断学习和积累经验的过程，要保持学习的热情和耐心，不断提升自己的科研能力和学术素养。与同行交流和合作，分享经验和心得，共同进步。

问题 7：您在教学上也获得了多项奖励，如厦门大学教学成果一等奖、厦门大学第十届青年教师教学技能比赛二等奖等，请问您在繁忙的科研之余，又是如何在教学上有所突破的？您的教学理念是什么？

作为一名高校教师，教学是一项重要的责任和使命，也是我们的初心和使命。我需要了解班上学生的学习风格、兴趣和能力，并根据这些信息设计个性化的教学方式。比如给信息学院学生上的《计算机导论》，我会注重具体技术层次的探索和分析。而给管理学院上的《人工智能》，我更注重基本原理层面的讲授。我通常倡导课堂上的互动与合作，通过提问、讨论和小组活动等方式激发学生的思维，培养他们的批判性思维和解决问题

的能力。我相信学生是课堂的主体，他们应该成为学习的主动者。我喜欢运用真实的案例，特别是在不同公司期间做的一些产品经历来说明理论知识的应用，让学生通过分析和讨论案例，加深对知识的理解，并培养他们的实践能力和问题解决能力。

问题 8：您曾与章毓晋老师合作出版了专著《基于子空间人脸识别》，可否分享一下您在出版此专著方面的一些故事或经验？

与我的博士导师章毓晋老师合作出版专著《基于子空间人脸识别》是我科研生涯中的一段宝贵经历。我们从最初的构想、内容规划，到具体章节的撰写和修改，每一个环节都需要我们共同的努力和配合。在这个过程中，在章老师全程指导下，我们团队协作分工。在撰写专著的过程中，我们注重对人脸识别领域的最新研究成果进行总结和归纳，力求为读者呈现一本既系统又创新的专著。我们不仅回顾了已有的研究成果和方法，还提出了一些新的理论和观点，为该领域的研究和实践提供了有益的参考和启发。专著的出版过程是一个漫长而又繁琐的过程，需要经历稿件的提交、审稿、修改、校对等多个环节。我们在这个过程中需要耐心和细心，不断地修改和完善稿件，确保专著的质量和准确性。最终，经过多次修改和审核，我们的专著顺利完成了出版。

问题 9：如果吐露研究工作者的心声，您最想说的是什么？

在科研的道路上，我们时常会面临挑战和困难，但也会有无尽的乐趣和成就感。每一次探索、每一次突破都是一次珍贵的经历，不断努力和坚持不懈，是我们前行的动力。在这个过程中，我们可能会经历失败和挫折，但不要轻易放弃，因为每一次挑战都是成长的机会。

责任编委 余烨 赵振兵

严严

厦门大学信息学院计算机系教授，博士生导师，计算机系教工党支部副书记。现为国际知名期刊 Neurocomputing（JCR 2 区）、Visual Computer（JCR 2 区）责任编委（Associate Editor）、计算机研究与发展（CCF 中文期刊 A 类）青年编委、中国图象图形学报（CCF 中文期刊 B 类）青年编委、IEEE 高级会员、CCF 高级会员、CCF 计算机视觉专委会委员、福建省计算机学会第九届理事会理事、CSIG 厦门中心执委、厦门市公安局警务战略特邀研究员、福建省和厦门市高层次人才（B 类）。

2004 年 07 月毕业于电子科技大学电子工程学院，获工学学士学位。2009 年 01 月毕业于清华大学电子工程系图像图形研究所（导师：章毓晋教授），获工学博士学位。2009 年 03 月至 2010 年 09 月，在日本东京 NOKIA（诺基亚）研发中心（NOKIA 全球六大研发中心之一）图像软件组（Imaging Software Team）工作，担任算法助理研究员。2010 年 09 月至 2011 年 05 月，在新加坡 PANASONIC（松下）研究院（PANASONIC 海外最大研究院）多媒体处理组（Media Processing Group）工作，担任项目负责人。2011 年 06 月至今工作于厦门大学信息学院计算机系，从事教学和科研工作。

曾担任第 6 届网络多媒体计算与服务会议（ICIMCS）本地主席，第 7 届 IEEE 未来多媒体技术会议（FMT）本地主席，第 6 届中国模式识别与计算机视觉大会（PRCV 2023）本地协调主席。还担任权威期刊如 IEEE T-PAMI、T-IP、T-NNLS、T-CSVT、T-ITS、PR、Neurocomputing、JVCI、计算机研究与发展、电子与信息学报、自动化学报等审稿专家等和国际权威会议如 ICCV、CVPR、ECCV、AAAI、ACM MM、ICIP、ICPR、ICIG 程序委员会委员。

委员好消息

✪ 2024 年 3 月 30 日，江苏省教育厅公示了 2023 年度江苏省高等学校科学技术研究成果奖拟获奖项目，CCF-CV 专委会副主任委员、南京信息工程大学**刘青山**主持完成的“高分辨率遥感影像高效获取与智能解译方法”拟授二等奖，CCF-CV 专委会委员、南京理工大学**李泽超**参与完成的“面向低质数据鲁棒性分析的多粒度建模方法研究”拟授三等奖。

✪ 2024 年 3 月 22 日，陕西省人民政府发布了关于 2023 年度陕西省科学技术奖励的决定，CCF-CV 专委会 5 位执行委员完成的 3 项成果获奖，分别是：西安交通大学**薛建儒**等完成的“智能驾驶的动态场景模式表征与预测理论及方法”、西安电子科技大学**邓成**、**高新波**（现重庆邮电大学）、上海交通大学**严骏驰**等完成的“多模态数据统一表征学习理论与方法研究”获自然科学一等奖，西安电子科技大学**苗启广**等完成的“多源遥感影像配准理论与关键技术”获自然科学二等奖。

✪ 2024 年 3 月 22 日，山西省人民政府发布了关于 2023 年度山西省科学技术奖励的决定，CCF-CV 专委会 2 位执行委员完成的 2 项成果获奖，分别是：太原理工大学**赵涓涓**等完成的“早期肺癌智能筛查的关键技术研究及应用”获技术发明二等奖，中北大学**秦品乐**等完成的“室外智能健身系统测练管一体化关键技术研究及应用”获科技进步二等奖。

✪ 2024 年 4 月 6 日，福建省人民政府发布了关于 2022 年度福建省省科学技术奖励的决定，CCF-CV 专委会 6 位执行委员完成的 5 项成果获奖，分别是：厦门大学**曲延云**等完成的“基于‘知识+层次’的视觉感知理解理论和方法”获自然科学二等奖，厦门大学**纪荣嵘**等完成的“城市大脑视觉数据高效感知与智能中台分析技术及其产业化”和北京交通大学**张淳杰**等完成的“病理 AI 辅助诊断平台关键技术研发与产业化”获科技进步二等奖，杭州电子科技大学**俞俊**等完成的“基于图论的多模态图像模式识别理论与应用”获自然科学三等奖，福州大学**牛玉贞**和**赵铁松**等完成的“跨行业异构数据智能分析关键技术与应用”获科技进步三等奖。

✪ 2024 年 4 月 9 日，云南省科学技术厅公布 2023 年度云南省科学技术奖拟奖项目名单，CCF 专委会执行委员、云南大学**陶大鹏**等完成的“基于多源信息融合与鲁棒性特征提取的行人重识别研究”获自然科学二等奖。

✪ 2024 年 4 月 26 日，《麻省理工科技评论》中国×DeepTech 发布“2023 年中国智能计算创新人物”入选者名单，CCF 专委会执行委员、北京师范大学**邬霞**和上海交通大学**严骏驰**入选。

✪ 2024 年 4 月 30 日，江西省科学技术奖励委员会办公室发布 2023 年度江西省科学技术奖拟推荐候选项目公告，CCF 专委会执行委员、江西财经大学**方玉明**等完成的“视觉信息自适应感知优化理论与方法”被推荐授予自然科学一等奖。

✪ 2024 年 5 月 1 日，上海市科学技术奖励管理办公室公示了 2023 年度上海市科学技术奖复评结果，CCF-CV 专委会 5 位执行委员的成果入选，他们是：复旦大学**姜育刚**等完成的“面向智能制造的跨域融合感知关键技术及应用”拟授技术发明一等奖，同济大学**赵才荣**、华中科技大学**白翔**等完成的“面向行人重识别的高效学习理论与方法”拟授自然科学二等奖，复旦大学**张文强**等完成的“中医面诊数字化、智能化识别的关键技术体系构建与应用技术体系构建与应用”拟授技术发明奖二

等奖。

✪ 2024年5月13日，山东省人民政府印发关于2023年度山东省科学技术奖励的决定，CCF-CV专委，10位执行委员的6项成果获奖：山东师范大学**朱磊**主持完成的“异构多源数据分析与融合方法研究”获自然科学二等奖，山东大学**聂礼强**（现哈尔滨工业大学）、华北电力大学**赵振兵**、**翟永杰**、山东大学**甘甜**等完成的“基于视觉大模型的输电线路精细化巡检关键技术及应用”获技术发明一等奖，山东大学**张伟**主持完成的“动态场景机器人高效决策与优化控制关键技术及应用”获技术发明奖二等奖，天津大学**刘安安**、中国海洋大学**黄磊**参与完成的“海洋大数据与智能计算平台技术研发及应用”获科技进步一等奖，山东财经大学**蹇木伟**主持完成的“面向复杂场景的公共安全智能信息处理关键技术及应用”、山东财经大学**崔超然**参与完成的“面向智慧法院的案件审判风险防控关键技术及应用”获科技进步二等奖。

✪ 2024年6月19日，ACM顶会SIGGRAPH 2024最佳论文揭晓，CCF-CV专委会常务委员、上海科技大学**虞晶怡**等完成的论文CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets和DressCode: Autoregressively Sewing and Generating Garments From Text Guidance获荣誉提名。

✪ 2024年6月24日，2023年度国家科学技术奖励结果揭晓，CCF-CV专委会执行委员、西安交通大学**孙剑**、**孟德宇**等完成的“弱观测成像反问题的L（1/2）理论与自适应正则化方法”获自然科学二等奖，CCF-CV专委会执行委员、天津大学**雷建军**等完成的“集成光场3D显示关键技术及应用”获技术发明二等奖。

责任编委 刘海波

基于缺失模态脑肿瘤自动分割开源代码

东北大学 贾同

脑肿瘤是致命的恶性肿瘤，从多模态 MRI 序列中其自动分割对临床评估和治疗至关重要。因此，许多研究广泛采用多模态 MRI 数据自动分割脑肿瘤，面对临床中常见的数据缺失，提出了多种应对策略。它们可分为三类。多模态融合方法，多模态融合方法通常忽略缺失模态数据中的有用信息，仅利用现有模态集特征；基于生成模型合成缺失的模态，在利用完整的模态进行模型优化；基于共享知识提取的方法学习潜在的特征空间。本文重点介绍后两种方法的研究成果。

1、Adversarial Co-training Network (ACN)

ACN 采用对抗性协同训练法，通过模态生成恢复缺失信息，实现模态间的互补学习。具体而言，模型包括两条学习路径：获取丰富模态信息的多模态路径和生成模态特殊表示的单模态路径。ACN 依靠三个核心子模块来实现：熵对抗学习模块（EnA）弥补多模态和单通道之间的领域差距；知识对抗学习模块（KnA），鼓励多模态和单通道学习中潜在表示对齐；模态互信息知识转移模块（MMI），通过变分信息最大化从不同特征层次恢复不完全模态的最相关知识，其结构图如图 1 所示。

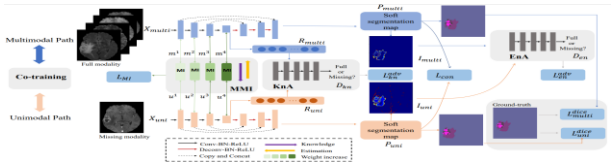


图 1 ACN 结构图

多模态训练的模型通常比单模态训练更可靠。因此，ACN 采用熵映射的原理作为无监督域自适应（UDA）任务的一致性度量，通过鼓励单模态学习的熵分布更接近多模态分布来最小化单模态学习的熵。然后引入一个全卷积网络作为鉴别器 D_{en}。

多模态训练模型因信息丰富而更可靠，但需对齐潜在空间特征分布。最小化两条路径的瓶颈特征之间的 KL 差异，容易干扰深层单模态路径的底层学习。因此使用一个知识对抗性模块来对齐这两条路径的高级表示。与 EnA 类似，生成器 G_{kn} 试图生成高级特征，以误导另一个鉴别器 D_{kn}。KnA 作为一种软对齐，鼓励单模态路径从完整的模式中学习 “缺失” 的知识。MMI 通过对每对熵和条件熵之间的互信息使用变分信息最大化来减少编码层中多模态学习知识的不确定性。

实验表明在所有模态缺失下，ACN 在多模态 BraTS2018 数据集上明显优于现有的方法。这种 “一站式” 方法对临床应用有很大价值，也可以推广到其他不完整的数据领域。更多有关 ACN 的详细内容可参考论文 “ACN: Adversarial Co-training Network for BrainTumor Segmentation with Missing Modalities”

论文地址：

https://arxiv.org/pdf/2106.14591.pdf

代码地址：

https://github.com/Wangyixinxin/ACN

2、Dual Disentanglement Network (D²-Net)

基于共享知识提取法可建立多模态关联，但存在训练不稳定和特征分离不理想的问题。在此工作中，D²-Net 提出了一种基于空间和频率域对比学习的方案，以明确和稳定地分解模式特定信息，并引入了一种新的引导密集知识蒸馏机制，其包括用于模态特异性信息解耦的模态解纠缠模块（MD）和用于肿瘤特异性知识分离的肿瘤区域解纠缠模块（TD），如图 2 所示。

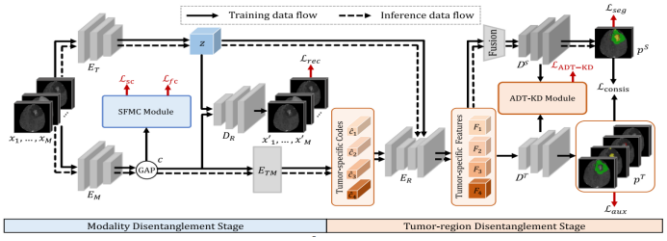


图 2 D²-Net 结构图

在 MD 阶段，D²-Net 以多种模式的 MRI 图像 x 作为输入，通过肿瘤编码器 E_T 获得融合的肿瘤特征 z。同时，在设计的空间频率联合模态对比（SFMC）学习方案的约束下，将每个模态图像分别输入与全局平均池化（GAP）层对应的模态编码器 E_M，生成特定模态的编码 c。然后通过肿瘤模态投影网络 E_{TM} 将解纠缠的模态特异性码 c 转移到所需的肿瘤特异性码 $\hat{c}$ 上，以发现潜在空间中模态与肿瘤区域特征之间的相关性，SFMC 模块结构如图 3 所示。

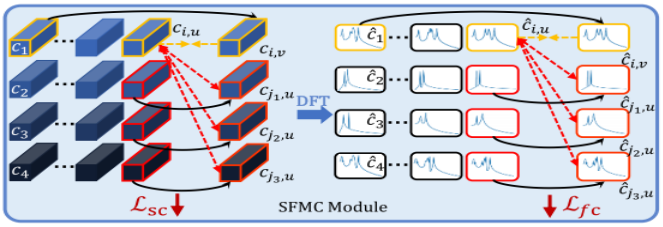


图 3 SFMC 模块结构图

在 TD 阶段，内容重构网络 E_R 转移肿瘤特征 z 和肿瘤特异性码 $\hat{c}$ 为肿瘤特异性特征 F。肿瘤区域特征 F 通过解耦二进制解码器 D^T，然后提高整体特征通过一个新颖的亲和力引导密集肿瘤区域知识蒸馏（ADT-KD）机制与整体多类肿瘤区域解码器 D^S 和解开二进制解码器 D^T。最后，利用整体肿瘤区域的学生 D^S 进行脑肿瘤分割。ADT-KD 模块结构如图 4 所示。

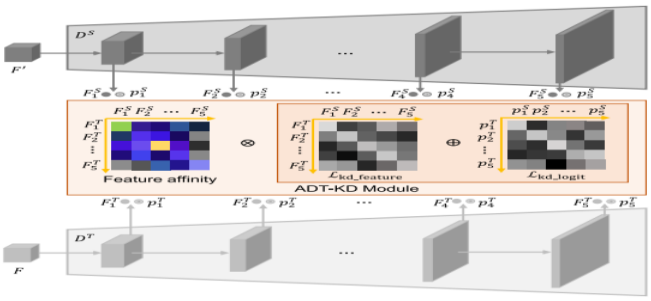


图 4 ADT-KD 模块结构图

该文在 BraTS2018 数据集上进行的大量实验表明，与最先进的方法相比，D2-Net 模型在缺失模式的脑肿瘤分割方面的能力和鲁棒性。

更多有关 D²-Net 的详细内容可参考发布该方法的论文“D2-Net: Dual Disentanglement Network for Brain Tumor Segmentation With Missing Modalities”。

论文地址：
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9775681

代码地址：
https://github.com/CityU-AIM-Group/D2Net

责任编委 李策 王田

贾 同

东北大学信息科学与工程学院教授、博士生导师，入选国家重大人才工程，人工智能系科研主任，智能感知与机器人研究所所长，辽宁省智能科学与智能系统重点实验室副主任。研究方向为计算机视觉、模式识别、图像处理和深度学习等。电子邮箱：jiatong@ise.neu.edu.cn

医学图像分割数据集

兰州理工大学 李策 张建伟

医学图像分割算法作为一种重要的医学图像处理技术，其旨在将医学图像中正常组织器官及病变区域检测并区分出来，并从分割的区域分析病变在图像上的表现特征，使临床诊断的准确性和可靠性得到有效提高，为患者的病情判断、疾病的临床诊疗和预后管理等提供可靠的依据。

医学分割算法的成功取决于专家提供的具有相应标签的高质量图像数据的。本文重点介绍了医学图像分割领域一些常见的公开数据集。

1、MSD 数据集

MSD 数据集(Medical Segmentation Decathlon, MSD)是一个公开的医学图像分割数据库。由十个不同任务的医学图像数据集组成，分别是：大脑肿瘤分割、心脏分割、肝脏分割、海马体分割、前列腺分割、肺部分割、胰腺分割、肝血管分割、脾脏分割和结肠分割任务。每个任务包含原始 CT 或 MRI 图像及其对应的手动分割标注，可用于开发和评估医学图像分割算法。数据以 NIfTI 格式存储。图 1 所示为 MSD 数据集中的部分图像数据示例。

数据集下载地址：http://medicaldecathlon.com/

相关论文链接：(1) The medical segmentation decathlon

https://www.nature.com/articles/s41467-022-30695-9

(2) A large annotated medical image dataset for the development and evaluation of segmentation algorithms.https://arxiv.org/pdf/1902.09063

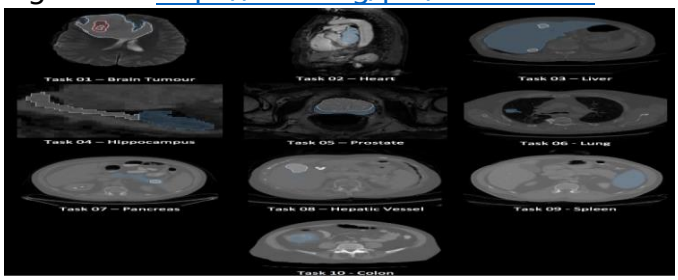


图 1 MSD 数据集中部分图像示例

2、BTCV 数据集

BTCV 数据集(Multi-Atlas Labeling Beyond The Cranial Vault, BTCV)是由范德堡大学医学中心(Vanderbilt University Medical Center)提供的一个用于腹部 CT 数据多器官分割的公开数据集。共包含 50 份腹部 CT 扫描数据，这些扫描源自转移性肝癌患者或术后腹壁疝患者，其中训练集有 30 例，测试集有 20 例，该数据集中提供了腹部 13 个解剖结构的分割标注，包括：肝脏、脾脏、胰腺、左肾、右肾、胆囊、腹主动脉、下腔静脉、食道、胃、十二指肠、胰管和脾门。
数据集下载地址：https://www.synapse.org/#!Synap

3、TotalSegmentator 数据集

TotalSegmentator 数据集来自巴塞尔大学医院放

射学和核医学诊所，是目前三维医学图像分割领域的最大公开数据集。它有两版数据，第一版数据于 2022 年 7 月公开，而后官方在 2023 年 9 月对数据集进行了较大的更新，增加了图像数量和标注类别数。图像总数从 1204 例增加到 1228 例(仅增加测试集数量)，类别数从 104 类增加到 117 类。图 2 所示为 TotalSegmentator 数据集中 104 个解剖结构概述。

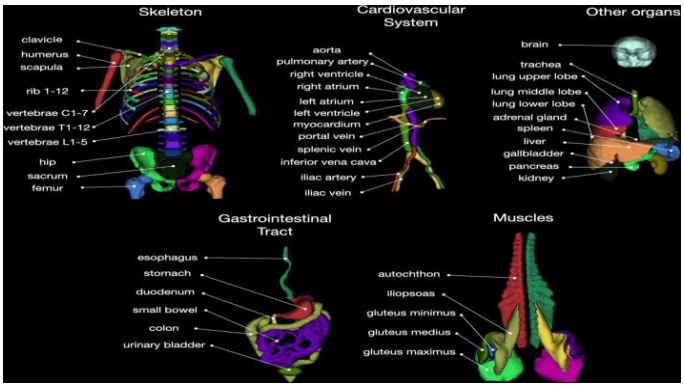


图 2 TotalSegmentator

数据集下载地址：
https://zenodo.org/records/6802614

相关论文链接： TotalSegmentator: robust segmentation of 104 anatomical structures in CT images, https://arxiv.org/abs/2208.05868

4、ISIC 2017 数据集

ISIC 2017 数据集(International Skin Imaging Collaboration, ISIC)是一个用于皮肤病变和皮肤癌诊断的皮肤镜图像数据集，包含 2000 张训练图像，150 张验证图像和 600 张测试图像。

数据集下载地址：
https://challenge.isicarchive.com/data/#2017

5、FLARE 数据集

FLARE 数据集(Fast, Low-resource, and Accurate Organ and Pan-cancer Segmentation in Abdomen CT, FLARE)是一个精确分割腹部多器官和肿瘤的数据集。它从 2021 年开始以挑战赛的形式发布，共有三个版本。21 版本仅有肝、脾、胰、肾四个腹部器官，包含 361 例训练集，50 例验证集以及 100 例测试集。22 版本包含腹部 13 个类别的器官分割，包含 2300 张 CT 数据，训练集只有 50 例有标注，2000 例无标注，另外还有 50 例验证数据及 200 例测试数据。从 22 版本开始，FLARE 的任务转向如何挖掘无标注数据的价值。23 版本同样包含腹部 13 个类别器官分割和肿瘤分割，训练集共 4000 个，由 2200 例部分标注和 1800 例无标注数据组成。另外提供 100 例在线验证数据和 400 例测试集数据。

数据集下载地址： https://codalab.lisn.upsaclay.fr/competitions/12239#learn_the_details-dataset

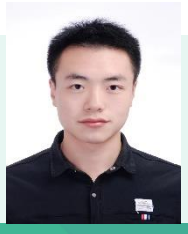
https://flare22.grand-challenge.org/Dataset/

https://flare.grand-challenge.org/Data/

责任编委 樊鑫 王田

李 策

教授，博士生导师，兰州理工大学电气工程与信息工程学院从事教育与科研工作，任网络与信息中心主任。研究方向为计算机视觉、医学影像分析，智能机器人等。



张建伟

兰州理工大学电气工程与信息工程学院 硕士研究生，研究方向为医学影像分析、计算机视觉等。

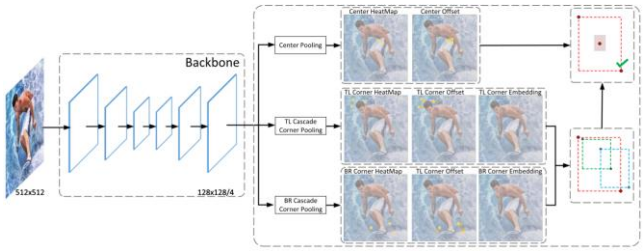
好 文 推 荐

来自中科院大学，英国牛津大学和深圳华为有限公司的研究人员将最新完成的论文成果“CenterNet++ for Object Detection”发表在 IEEE TPAMI 2024。

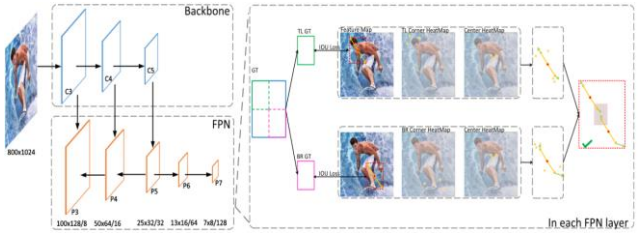
论文：Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, (IEEE Fellow), and Qi Tian (IEEE Fellow). CenterNet++ for Object Detection, IEEE TPAMI, 46 (5): 3509-3521, 2024

现阶段，目标检测有两种主流方法，分别是：自顶向下的目标检测方法和自底向上的目标检测方法。许多研究人员认为自底向上的方法耗时长，且会引入较多的假阳性，而自顶向下的方法因在实践中通常获得更好的检测效果，因此自顶向下的方法逐渐成为主流方法。自顶向下方法将每个目标建模为先验点（a prior point）或预定义的锚框（a predefined anchor box），并预测与边界框相对应的偏移量。自顶向下的方法能够感知整个目标，且极大简化了生成边界框的后处理步骤。然而，这类方法往往无法很好预测形状奇特的目标（例如，具有大长宽比的目标）。对应地，自底向上方法则可以很好预测具有任意几何形状的目标，因此该类方法具有更好的召回性能。由于现有自底向上方法易出现假阳性，导致该类方法无法很好地表示目标。

该研究团队在其最新论文成果中证明了自底向上的方法相较于自顶向下的方法更具检测竞争力，同时具备更优秀的召回性能。首先，团队研究人员提出一种自底向上的目标检测方法，名为 CenterNet++。CenterNet将每个目标的检测定义为一个三元组关键点的检测，从而可以定位具有任意几何形状的目标，同时可感知目标的全局信息。



(a) 单分辨率检测框架



(b) 多分辨率检测框架

图 1 所提方法 CenterNet++的结构流程图

接下来，团队研究人员设计了两个框架来适应不同结构的网络，提高了所提方法的网络泛化能力。如图(1)所示，这两个框架分别是单分辨率检测框架和多分辨率检测框架。前者应用于单分辨率特征检测，后者应用于多分辨率特征（例如：特征金字塔）检测。最后，该团队将所提方法与目前主流的自顶向下和自底向上的检测方法相比较，CenterNet++均取得优异的检测结果，同时获得优秀的召回性能。

责任编委 贾同 樊鑫

好 文 推 荐

来自东北大学，北京大学和长沙海信智能系统研究的研究人员将最新完成的论文成果“FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection”发表在 IEEE TPAMI 2024。

论文：Shuailei Ma, Yuefeng Wang, Shanze Wang, and Ying Wei. FGAHOI: Fine-Grained Anchors for Human-Object Interaction Detection, IEEE TPAMI, 46 (4): 2415-2429, 2024

人-物交互（Human-Object Interaction，HOI）是计算机视觉领域中一个重要问题，旨在定位（人，物）对并识别两者之间的交互关系。相较于单个目标实例，HOI 问题具有更大的空间、尺度和任务跨度，使其检测更容易受到背景噪声的影响。为了缓解背景噪声对 HOI 检测的干扰，研究人员往往首先需要利用输入图像的信息生成细粒度的锚点，然后利用这些锚点指导 HOI 实例的检测。然而，这类方法面临以下挑战：（1）现阶段尚未有可以从具有复杂背景的图像中提取关键特征的有效方法。（2）目前研究人员难以将提取的特征与查询嵌入进行语义对齐。

针对上述难题，该团队研究人员提出一种基于 Transformer 的端到端人-物检测框架框架，名为 FGAHOI。如图（1）所示，FGAHOI 主要由 3 部分组成，分别是：多尺度采样模块（Multi-Scale Sampling，MSS），分层空间感知融合模块（Hierarchical Spatial-Aware Mergin，HSAM）和任务感知融合机制（Task-Aware Merging，TAM）。简单而言，MSS 旨在从背景噪声中提取人、物和交互区域的特征，用于不同尺度的 HOI 实例。HSAM 和 TAM 依次从层次空间和任务角度对提取的特征和查询嵌入进行语义对齐和合并。此外，该团队研究人员还设计一种新的训练策略，名为分阶段训练策略，以减少所提 FGAHOI 因任务过于复杂而带来的训练压力。为了多角度衡量 HOI，该团队研究人员进一步提出了一个新的数据集 HOI-SDC。HOI-SDC 从（人，物）对分布不均匀和（人，物）对远距离视觉建模两方面评估 HOI 任务的性能。最后，该团队研究人员在所提数据集 HICO-DET 和 HOI-SDC 以及现存的人-物交互数据集 V-COCO 上进行了大量对比实验，验证了所提方法 FGAHOI 的优越性。同时，对 FGAHOI 的 3 个组成部分进行了细致的性能分析实验。

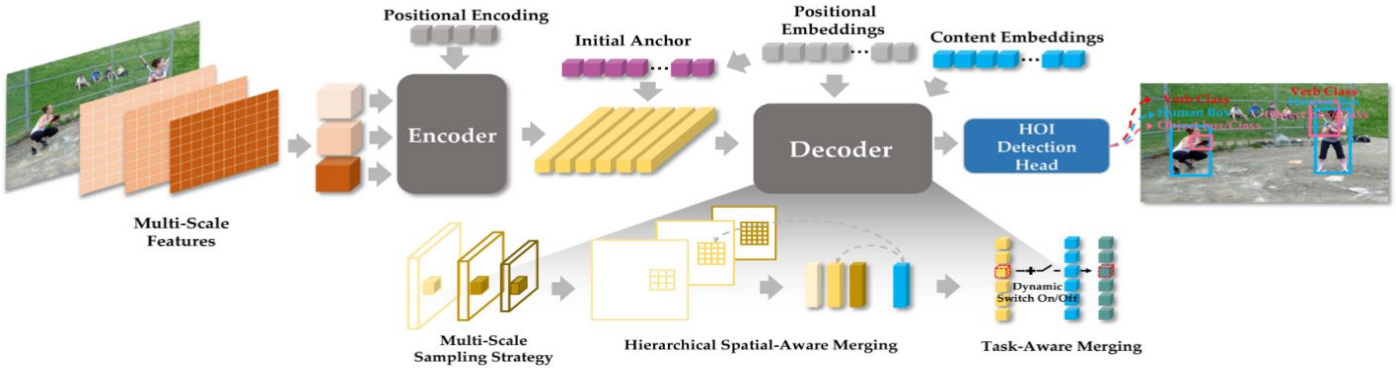


图 1 所提方法 CenterNet++的结构流程图

责任编委 李策 贾同

好 文 推 荐

南京大学的”IOMatch: Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization.” 发表于 ICCV 2023 会议，获得了三位审稿人一致的”Strong Accept”评价，并被推荐为口头报告论文。代码和模型已开源。（https://github.com/nukezil/IOMatch）

论文：Zekun Li, Lei Qi, Yinghuan Shi, Yang Gao. IOMatch: Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization. ICCV, 15870–15879, 2023.

基于伪标注和一致性正则化的深度半监督学习方法已经取得了巨大的成功。然而，现有大多数方法仍然依赖于一个关键的闭集假设：训练过程中的有标注样本和无标注样本的类别空间完全相同。在开集场景下，当有标注样本和无标注样本存在类别差异时，现有的深度半监督学习方法可能遭受严重的性能损失。对此，已有的开集半监督学习方法主要采取一种直觉化的“检测-排除”范式，即先将无标注数据中的未知类样本检测出来，再将它们排除在半监督训练之外。然而，当有标注数据量极少时，预训练得到的未知类检测器将是非常不可靠的，它将会错误地把相当比例的已知类无标注样本判别为未知类样本并抛弃。这样的错误不断积累，直至最后绝大部分无标注数据都将被排除在半监督训练之外，从而造成更为显著的性能下降。事实上，即使是专门针对分布外检测问题设计的先进方法，在标注信息极少的条件下也难以取得令人满意的检测性能。因此，本文没有沿用“检测-排除”的主流范式，而是另辟蹊径，设法构建一个无需依赖准确已知类/未知类判别的开集半监督学习框架 IOMatch。

IOMatch 的核心思路是：对所有的开集无标注数据（同时包括已知类和未知类样本），生成并利用形式统一的开集伪标签。具体地，假设已知类别的数目为 K，那么相应开集伪标签向量的维度应当为 K+1 ，其中所有的未知类样本都被视作同一个"Unknown"类别。本文在 FixMatch 框架（仅包含一个闭集分类器）的基础上，额外引入一个多重二分类器以及一个开集分类器，如图 1 所示。本文提出了一种按类加权的方式的来融合闭集分类器与多重二分类器的预测结果，所得开集伪标签作为监督信号用于训练开集分类器。通过这样的方式，IOMatch 能够更充分地同时利用无标注数据中的已知类和未知类样本，在训练过程中逐渐增强已知类分类和未知类检测的能力。在多个开集半监督学习测试基准上，IOMatch 取得了显著优于当前先进方法的性能，同时展现出高度的简洁性。

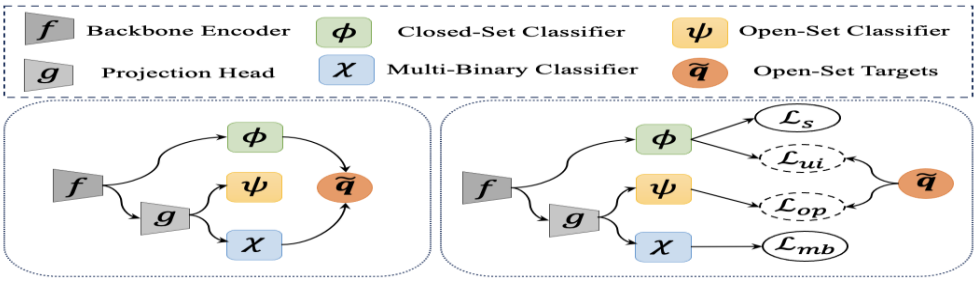


图 1 开集半监督学习框架 IOMatch 示意图

责任编委 王田 樊鑫

征文通知

1 会议征文

计算机视觉领域相关国内外会议的征文通知如表 1 所示。同时，可继续关注每个会议举办的 workshop 或 special session。

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 IEEE Journal of Biomedical and Health Informatics, Pattern Recognition, Decision Support Systems 和 Computer Vision and Image Understanding。

3 会议简介

中国模式识别与计算机视觉学术会议 PRCV (Chinese Conference on Pattern Recognition and Computer Vision)，由中国计算机学会（CCF）、中国自动化学会（CAA）、中国图象图形学学会（CSIG）和中国人工智能学会（CAAI）联合主办，定位国内顶级的模式识别和计算机视觉领域学术盛会。

第七届 PRCV 将于 2024 年 10 月 18 日至 10 月 20 日在乌鲁木齐举办，由新疆大学承办。本届会议旨在汇聚国际国内模式识别和计算机视觉领域的广大科研工作者及工业界同行，分享最新理论研究进展和技术研发成果。通过此次会议，能加强本领域学术界和企业界进行深入的“产学研”交流与合作，从而进一步促进模式识别与计算机视觉领域的协同创新。

责任编辑：刘帅奇

表 1 计算机视觉领域相关国内外会议

会议名称	会议时间	会议地点	截稿日期	会议网站
ACCV 2024	2024.12.08-12	Hanoi, Vietnam	2024.07.01	https://accv2024.org/
ICTAI 2024	2024.10.28-30	Herndon, Virginia, USA	2024.07.01	http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=179555
AAAI 2025	2025.02.25-03.04	Philadelphia, Pennsylvani, USA	2024.08.07	https://aaai.org/aaai-conference/save-the-date-aaai-25/

表 2 计算机视觉领域相关国内外期刊专刊

期刊名称	专刊题目	投稿网址	截稿日期
CVIU	Advanced Computational Imaging and Photography Measurement	https://www.sciencedirect.com/journal/computer-vision-and-image-understanding	2024.09.15
PR	Conformal Prediction and Distribution-Free Uncertainty Quantification	https://www.sciencedirect.com/journal/pattern-recognition/about/call-for-papers#conformal-prediction-and-distribution-free-uncertainty-quantification	2024.09.30
DSS	Empowering Bright Internet and Bright Artificial Intelligence	https://www.sciencedirect.com/journal/decision-support-systems/about/call-for-papers#empowering-bright-internet-and-bright-artificial-intelligence-ai	2024.07.31
JBHI	Multimodal Approaches in Neuroimaging with Explainable and Responsible AI	https://www.embs.org/jbhi/wp-content/uploads/sites/18/2024/05/Multimodal-Approaches-in-Neuroimaging-with-Explainable-and-Responsible-AI.pdf	2024.08.31

COMPUTER VISION NEWSLETTER

02 2024

总第 40 期

计算机视觉专委会简报

CCF 计算机视觉专委会