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

03 2022

总第 33 期

中国计算机学会 CCF

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会 简报

2022 年第 03 期

总第 33 期

CCF 计算机视觉专委会

主　办　CCF 计算机视觉专业委员会

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 114 期 燕山大学

2022 年 6 月 18 日上午，由中国计算机学会计算机视觉专委会（CCF-CV）主办、燕山大学承办的 CCF-CV 走进高校系列报告会第 114 期活动以线上方式成功举办。本次活动邀请了北京交通大学赵耀教授、北京大学彭宇新教授、中国科学技术大学张勇东教授做特邀报告。燕山大学信息科学与工程学院（软件学院）副院长练秋生教授和燕山大学信息与通信工程学科负责人、中国计算机学会计算机视觉专委会执行委员顾广华教授担任本次活动的执行主席。练秋生教授主持本次学术报告会。

本次报告会采取 CCF 计算机视觉专委会 B 站官方账号直播+腾讯会议的方式，B 站人气峰值接近 1000，腾讯会议燕山大学教师参加人数近 80。最后，燕山大学信息科学与工程学院(软件学院)院长齐跃峰教授进行了活动总结，首先感谢了三位专家的精彩报告与学术交流分享，同时感谢了线上听众的热情参与和高质量提问，最后再次感谢 CCF-CV 专委会和学校对信息科学与工程学院承办本次活动的大力支持！祝贺本次活动取得了圆满成功！

第 115 期 北京工业大学

2022 年 7 月 15 日，由中国计算机学会计算机视觉专委会主办、北京工业大学承办的 CCF-CV 走进高校系列报告会活动，在 CCF 计算机视觉专委会 B 站官方账号成功举办。本次活动邀请了北京大学林宙辰教授、西安电子科技大学邓成教授、中国科学院计算技术研究所王瑞平研究员、江西财经大学方玉明教授以及上海交通大学马超副教授等五位计算机视觉领域专家学者做特邀报告。北京工业大学信息学部主任李晓理出席并致欢迎辞。北京工业大学信息学部胡永利教授、简萌副教授和王博岳副教授担任本次会议的执行主席。

最后，活动执行主席、北京工业大学胡永利教授对本次活动进行总结。首先感谢了五位报告专家分享了团队的最新研究成果和思路，也感谢参会的老师和同学的细心聆听。此外，各位报告专家站在学科前沿、不断学习、不断探索，是我们学习的榜样！最后，感谢中国计算机学会(CCF) 计算机视觉专委会、北京人工智能研究院以及北京工业大学信息学部给予本次会议的大力支持！

责任编委 毋立芳

RACV2022 计算机视觉前沿进展研讨会成功召开

2022 年 8 月 10-11 日，中国计算机学会计算机视觉专委会（CCF-CV）年度学术研讨会 RACV（Recent Advances on Computer Vision）在大连成功召开。RACV 定位为国内计算机视觉领域的小规模精品研讨会，通过定向邀请方式汇集领域专家，深度研讨计算机视觉领域中的若干核心问题并形成进展报告。研讨会试图通过务实、开放与平等的对话与讨论，深入发掘相关研究领域潜在的问题，为广大的科研人员提供观察问题的新视角与新观点。

本次会议开幕式由专委会副主任、上海科技大学虞晶怡教授主持，中国计算机学会秘书长唐卫清研究员和大连理工大学卢湖川教授致开幕辞。根据专委常委会前期的讨论票选，本次会议设置了 3 项研讨主题。每项主题首先由特邀嘉宾们进行主题发言，之后所有与会人员进行自由讨论。

10 日上午首先进行了主题一“视觉基础模型”的研讨。该主题由专委会常委、百度王井东博士、专委会常委、南开大学程明明教授、南开大学侯淇彬副教授负责组织，邀请了百度王井东博士、华中科技大学王兴刚副教授、华为谢凌曦博士、清华大学黄高副教授 4 位嘉宾进行主题发言。近几年，视觉基础模型吸引了大量学者的注意，与会嘉宾围绕掩码图像建模、自监督表征预训练、语言对视觉基础模型的作用等问题进行了精彩的讨论和观点分享。

10 日下午进行了主题二“三维重建和沉浸式渲染”的研讨。该主题由专委会副主任、上海科技大学虞晶怡教授、专委会常委、航天宏图王涛博士、上海科技大学李实英副研究员 3 位委员负责组织，邀请了中科院自动化所申抒含研究员、上海科技大学许岚助理教授、中科院计算所高林副研究员、宾夕法尼亚大学刘玲洁助理教授 4 位嘉宾进行主题发言。嘉宾们围绕神经网络渲染和生成技术、三维重建和沉浸式渲染的学术前沿和发展方向、元宇宙背景下的商业前景和潜在风险等议题展开了深入探讨，合合信息丁凯博士、郭丰俊博士发表观点并参与深入讨论。

11 日上午进行了主题三“具身视觉”的研讨。该主题由上海交通大学卢策吾教授、专委会常委、清华大学鲁继文副教授、北京大学王鹤助理教授负责组织，邀请了上海交通大学卢策吾教授、清华大学刘华平教授、中山大学郑伟诗教授、北京航空航天大学刘偲教授 4 位嘉宾进行主题发言。嘉宾们围绕具身智能的核心科学问题及在计算机视觉领域的角色、技术路线、评价标准、如何推动智能机器人落地等议题展开了深入探讨。

最后，研讨会闭幕式由专委会副主任、南京信息工程大学刘青山教授主持。中国计算机学会唐卫清秘书长对活动做了点评并提出改进建议，专委会主任、北京大学查红彬教授对活动做了总结并代表专委感谢承办方、赞助方和参会人员。本次研讨会在两天时间内深入探讨了本领域最前沿研究问题，主题发言视角广阔，自由讨论热情激烈，参会嘉宾们纷纷表示本次会议内容丰富，收获良多。按照计划，组委会后续将整理相关主题的发言与讨论文稿，形成观点性文档进行发布，把讨论从线下延伸到线上，欢迎更多专家学者积极参与。本次研讨会由大连理工大学卢湖川教授团队承办，合合信息提供独家赞助。

责任编委 杨巨峰

CCF-CV 常务委员会 2022 年度第一次工作会议顺利召开

中国计算机学会计算机视觉专委会（CCF-CV）常务委员会 2022 年度第一次工作会议于 6 月 29 日线上顺利召开！

本次常委会工作会议由专委会主任查红彬教授主持，常委会委员参会，秘书处全体成员列席。

首先，查红彬主任讲话。查主任充分肯定了专委会过去半年取得的工作成果，尤其肯定了疫情期间采用线上方式持续推进专委工作的努力，并对后续工作提出了切实可行的发展建议。

随后，专委会党小组组长、副主任刘青山教授组织了党小组学习。

接下来，常委会委员听取了专委会上半年的工作汇报及下半年活动规划，并针对专委发展相关议题展开了热烈讨论，逐一形成了具体可行的指导性建议，为专委会后续发展明确了方向和重点。专委发展相关议题来自专委会委员们的建言献策，秘书处进行分类汇总，涉及专委会学术活动、组织建设、领域发展等方面。

最后，查红彬主任作了总结发言。会议在紧张而有序的热烈讨论氛围中结束。本次常委会工作会议系统地梳理了专委会各项工作的进展，明确了专委会发展的工作重点，特别是对专委开创的各项品牌学术活动的继续创新突破指明了方向。

责任编委 黄岩

中国计算机学会计算机视觉专委会（CCF-CV）

2022 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，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。专委会成立八年以来，已经发展执行委员 354 人，在 CCF 专委评估中分别获得“特色活动奖”、“综合进步奖”、“优秀专委奖”、“年度特别奖”等 6 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 2022 年计算机视觉专委会的执行委员增选工作。

一、申请时间

2022 年 6 月 25 日— 2022 年 10 月 15 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“2022 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、特别说明

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 “2022 年中国模式识别与计算机视觉大会” （http://www.prcv.cn）期间（2022 年 11 月 4 日-11 月 7 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”**）。

责任编委 任传贤

专题综述

跨模态医学影像合成研究与展望

上海科技大学 潘永生 西北工业大学 崔恒飞 夏勇

一、引言

医学影像能够显示身体部位和器官的信息，在疾病诊断、治疗和预后预测中发挥着重要的作用。常见医学影像包括磁共振影像(MRI)、计算机断层成像(CT)、正电子发射成像(PET)、X 光平片、光学成像等。由于成像原理的差异，各种影像获取的信息有所不同，例如 MRI 能提供有关软组织的信息，CT 主要用于成像高电子密度组织(如骨骼)，但也可以提供一定程度的软组织对比度，PET 则使用放射性示踪剂成像特定的生物学功能。同时，根据成像参数和所用试剂的差异，同一种医学影像又有不同的子类型，比如 MRI 包括 T1 加权序列、T2 加权序列等，PET 包括 FDG-PET、Aβ-PET 等。图 1 给出了一些常见的医学影像。

同时包含不同类型或者子类型的医学影像则被称为多模态影像[1]。由于不同影像模态存在一定互补性，多模态影像在临床应用中通常能够比单模态影像提供更多信息。然而，多模态医学影像的获取会遇到采集时间长、费用高、可能增加辐射剂量等困难。因此，人们期待能够使用图像处理技术进行跨模态医学影像合成，即使用某一种(或一些)模态的医学影像去生成另一种(或一些)模态的医学影像[2]。

跨模态医学影像合成虽然能为多模态影像诊断带来便利，但也存在一些技术挑战，例如临床失效问题，即合成影像和真实影像在诊断性能上具有明显的差异[3]。这是因为，各类影像模态的成像原理不同，目标模态能采集的某些信息在源模态影像中并不存在，导致合成的目标模态影像依然不具有这些信息。同时，合成模型受

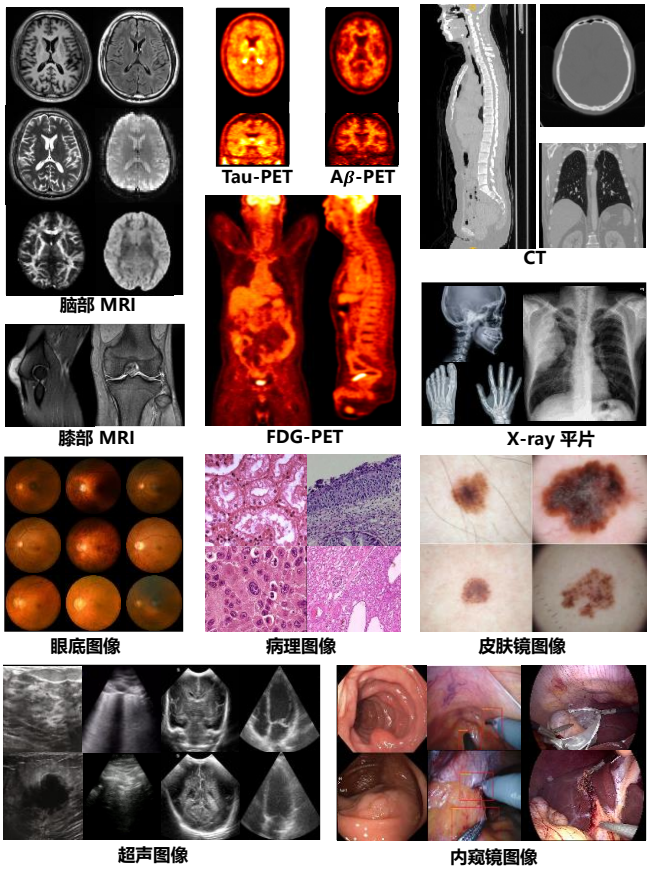


图 1　多种模态的医学影像

到自身表示能力的限制，会产生一定信息损失，特别当其训练过程受到约束条件的偏置化引导时，将在合成影像中产生与偏置相关的信息损失。当前，研究者们大多从模型本身入手，通过提高模型的表示能力或者设计针对具体任务的约束条件来提高合成影像的质量，所开发的跨模态医学影像合成技术已应用于影像采集、重建、配准、分割、检测、诊断等环节，给许多问题带来了新的解决思路和方法。

二、跨模态影像合成技术

2010 年以来，跨模态影像合成受到研究者越来越多的关注，产生了许多合成方法。本文将这些方法大致分为三类，包括传统合成方法、基于深度学习的合成方法和任务驱动的合成方法。

2.1 问题陈述

假设$\mathcal{X}=\{X_1,\ldots,X_s\}\sim\mathfrak{M}_X$ 是采集的S张源模态($\mathfrak{M}_X$)影像,$\mathcal{Y}=\{Y_1,\ldots,Y_T\}\sim\mathfrak{M}_Y$是采集的$T$张目标模态($\mathfrak{M}_Y$)影像。跨模态影像合成假设存在一种映射$\mathbb{G}:\mathcal{X}\to\mathcal{Y}$,

$$\forall\ X\sim\mathfrak{M}_X,\exists\ Y\sim\mathfrak{M}_Y\ \ s.t.\,\mathbb{G}(X)=Y.$$

并且，映射$\mathbb{G}$可以通过某种优化方法由给定的数据$\mathcal{X}$和$\mathcal{Y}$近似地估计出来，即

$$\widehat{\mathbb{G}}=\arg\min_{\mathbb{G}}\mathfrak{D}(X,Y;\mathbb{G}),$$
$$s.t.X\in\mathcal{X},Y\in\mathcal{Y}.$$

其中，$\mathfrak{D}$是反映约束条件的优化目标。目前，影像合成技术均围绕设计映射模型$\mathbb{G}$和优化目标$\mathfrak{D}$而展开，两者相辅相成。典型的映射模型有字典学习、随机森林、卷积网络、编解码网络等，常见的优化目标有平均绝对误差、结构相似性、对抗损失，特征一致性等。

2.2. 传统跨模态影像合成方法

这类方法通常将影像划分成多个小块，并将每个块编码成一个表示向量，通过建立不同模态的配对的块表示向量之间的映射，再根据源模态块的编码产生对应的目标模态块。主要关注表示向量的设计和映射模型的建立，模型的求解过程以类似“数据检索”的方式进行，优化目标通常使用平均均方误差等容易计算的指标。这类方法包括字典学习随机森林等。

基于字典学习的方法[4]假设每个模态存在一个字典，每个图像块均可由字典中元素的稀疏表示得到，不同模态对应的图像块具有相同的字典编码。进行跨模态影像合成时，为不同的模态设置统一的编码和不同的字典，并根据稀疏表示原理通过最小化联合重建误差来求解字典和编码。

基于随机森林的方法[5]将影像合成视为回归问题，假设目标模态块(或其中心点/中心区域)的值是源模态块的因变量，并且这种关系可以通过回归模型得到。这类方法需要首先使用其他方法编码每个图像块，因此非常受编码方式的影响。为了提高表示能力，随机森林使用的表示向量通常是由多个尺度的多种简单特征组合而成的，如空间位置、离散傅里叶系数、类 haar 特征、平均亮度等。

2.3. 基于深度学习的跨模态影像合成方法

随着深度学习的发展，跨模态影像合成研究已逐渐转移到深度学习框架中。从简单卷积神经网络(CNN)，到变分自编码网络(VAE)、U-Net、生成对抗网络(GAN)等，深度学习的各种技术都在跨模态医学影像合成中有所应用。与传统方法相比，此类方法可以直接使用大规模的参数化模型以端到端的方式建立从源模态影像到目标模态影像的映射，并以数据驱动的方式自动提取图像(块)的表示特征，而不需要手工设计表示特征。由于其便于实现和性能优越，基于深度学习的跨模态影像合成技术目前已经占据了主导地位。

2.3.1 基于简单 CNN 的方法

基于简单 CNN 的方法早期常采用与传统方法类似的策略，首先将图像划分成一系列的小块，并使用每个源模态图像块预测目标模态图像块的中心值或中心区域的值。不同的是，这些方法通过堆叠多个卷积层来构建映射模型，并将特征提取和回归预测集成在一起，使用误差反向传播算法来优化模型参数。由于其数据驱动的学习方式，这些方法在训练样本充足的情况下能够达到比传统方法更好的性能。但是，这些方法的表示能力受到参数量和结构复杂性的限制(如使用的卷积层数量和卷积核大小)。因此，早期的 CNN 方法通常沿着增加模型复杂度的方向发展。Rongjian Li 等只使用了两层卷积来建立 MRI 和 PET 之间的映射[6]，Dong Nie 等使用四层卷积来建立 MRI 和 CT 之间的映射[7]，而 Lei Xiang 等则进一步串联了 3 个四层的卷积网络(共计 12 层)来建立 T1-MRI 和低剂量 PET 到高剂量 PET 的映射[8]。然而，增加卷积层的个数在增强表示能力的同时也会增加计算复杂度；同时，由于依然需要逐块甚至逐像素计算目标值，这些方法的计算效率并不比传统方法更

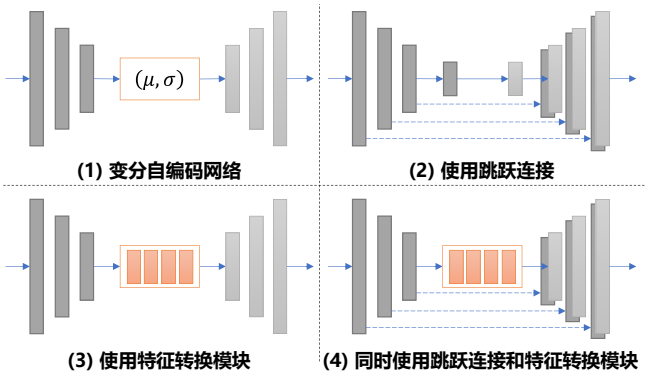


图 2　编解码网络及其变型

具优势。

2.3.2 基于编解码网络的方法

这类方法假设源模态和目标模态的影像在某一隐空间中存在共享的中间编码，因此其通常包含一个编码器和一个解码器，编码器将源模态图像(块)转换成中间编码，解码器将中间编码解码成目标模态图像(块)[9]。编码器通常采用多个带下采样的卷积网络来压缩源模态图像中的信息，解码器则采用多个带上采样的卷积网络来将压缩的信息恢复为目标模态图像。这类方法的优点是可以同时输出一个较大区域甚至整幅图像，不但避免了简单 CNN 方法中的逐像素计算，减少了计算代价，也有助于保留结构性信息，使合成图像具有更好的视觉效果。然而，这是一个“有损压缩”过程，存在信息丢失问题。例如，VAE 模型常采用均值和方差的组合作为中间编码，虽然保留了源模态图像中的统计信息，但同时抑制了个体化信息。有两种策略可以补偿损失的信息。一种是简化编码器和解码器的结构来减少信息损失，并在编解码之间增加额外的特征转换单元(如加入多个残差模块)来增强信息转换能力[2,10]。另一种是使用跳跃连接将编码器的中间特征同时作为解码器的输入来补偿缺失的信息[10, 11]。需要主要的是，一般的编解码结构可以适用于配对或非配对的图像，而带跳跃连接的结构则只有在配对的图像上才有良好的性能[12]。图 2 给出了用于影像生成的编解码网络及其变型的结构示意图。

2.3.3 生成对抗网络方法

简单 CNN 和编解码网络一般使用确定性的简单优化目标，如平均绝对误差(MAE)、均方误差(MSE)等，从而会引入确定性偏置，即合成网络的优化方向始终朝向优化目标，导致与优化目标相背的信息传输被抑制，使得合成的图像在被抑制的信息方面呈现出平均的效果，虽然在 MSE、PSNR、SSIM 等指标上表现优异，但看起来却非常模糊。为了克服这种确定性偏置的不利影响，出现了基于 GAN 的影像合成技术[1]。GAN 使用一个判别网络(判别器，$\mathbb{D}$)来指导生成网络(生成器，$\mathbb{G}$)的学习过程：通过交替训练生成器和判别器，使两个网络以相互竞争的方式同步提高各自的能力，最终(在理想情况下)达到纳什均衡。这个过程中，生成器逐渐生成具有真实图像特征的合成图像，判别器不断提高对合成图像的鉴别能力。假设将真实样本和合成样本输入判别器$\mathbb{D}$时的判别标签分别为 1 和 0，那么 GAN 中的$\mathbb{D}$和$\mathbb{G}$通过不断迭代如下两个优化过程求解：

$$\max_{\mathbb{D}} \mathbb{E}_{Y\sim\mathcal{Y}}\left[\log\left(\mathbb{D}(Y)\right)\right] + \mathbb{E}_{X\sim\mathcal{X}}\left[\log\left(1-\mathbb{D}\left(\mathbb{G}(X)\right)\right)\right],$$

$$\min_{\mathbb{G}} \mathbb{E}_{X\sim\mathcal{X}}\left[\log\left(1-\mathbb{D}\left(\mathbb{G}(X)\right)\right)\right].$$

判别器$\mathbb{D}$的存在，使得 GAN 可以直接优化似然度本身，而不是 MAE 等确定性目标的对数似然的下界。$\mathbb{D}$的不断更新相当于不断变换生成网络$\mathbb{G}$的优化目标，因此避免了引入确定性偏置。由于需要训练两个网络，GAN 在训练时相比于其他方法需要更多的计算时间和空间，但在应用时只需要运行生成网络，其时间复杂度和其他深度学习方法是接近的。

作为一种学习策略，任何可微的模型都可以用于构建判别器和生成器，但常用于影像合成的 GAN 通常使用图 2 中的一种结构作为生成器，而判别器则主要使用图 3-(1) 所示的多层卷积结构。训练 GAN 的理想状态是达到纳什均衡，这有时候可以用梯度下降法或其衍生算法做到，但大部分时候是做不到的。目前还没有方法确保能达到纳什均衡，所以相比于使用确定性优化目标的方法，GAN 的训练是不稳定的。为此，确定性优化目标不得不被重新考虑进来，如加入 MAE、SSIM 等。此外，也可以使用感知损失(Perceptual loss)来提高 GAN 的稳定性。但感知损失依赖额外的网络来提取中间特征，并且只能在不同模态的配对图像上计算。如果在固定参数的预训练网络上计算，感知损失实际上相当于非常多但有限的确定性优化目标的组合，仍然会引入一定程度

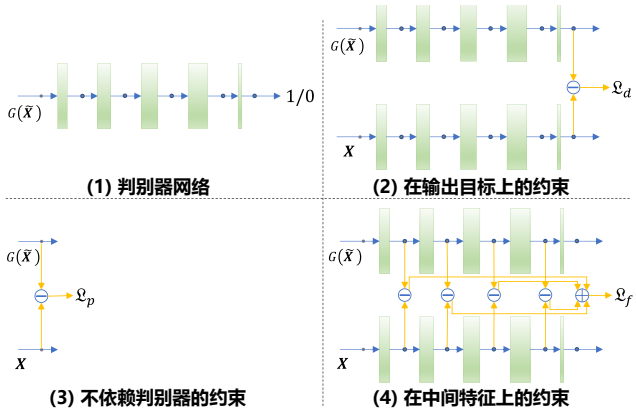


图 3　不同约束目标的对比示意图

的确定性偏置。此外，如果直接在判别网络上计算感知损失，也可以增强 GAN 的稳定性，此时多层感知损失也被称为特征匹配损失[13]。图 3 给出了不同约束目标的对比示意图，通过组合不同的约束目标可以得到 GAN 的不同变型。

三、面向任务的合成方法

许多跨模态影像合成技术并没有考虑具体问题的特点。因此，有的方法可能会产生“幻觉”，即倾向于合成训练数据中显著存在的图像模式，但该模式可能并不是下游任务所需要的[14]。为了解决该问题，需要在通用技术的基础上添加与任务相关的设计，形成对具体任务的偏置，从而使合成的图像保存更多有助于任务的信息，在具体任务上取得性能提升。前述感知损失就是一种这样的设计，其目的是为了平衡合成图像中的空间结构、形状与颜色、纹理等的保持与丢失。

在医学影像合成中，更多的是需要合成的影像对后续的诊断、分割、配准等任务有所帮助，这通常很难通过通用的合成方法实现，因为这些方法实际上主要包含与具体任务无关的偏置，这种无关偏置主导了模型的学习过程，使朝向具体任务的偏置被抑制了，进而产生“无用”的合成影像。这也是合成的医学影像在实际应用中很难得到认可的一个重要原因。为了提高合成的影像的“有用性”，针对具体问题设计专门的合成模型是一种有效的手段。因此，产生了一系列面向任务的影像合成方法[3, 15]。

3.1 面向任务的偏置

假设某种影像合成模型的驱动目标 $\mathfrak{D}$可以分解为与任务相关的部分$\mathfrak{D}_r$和与任务无关的部分$\mathfrak{D}_i$，即

$$\mathfrak{D}(X,Y;\mathbb{G}) = \mathfrak{D}_r(X,Y;\mathbb{G}) + \mathfrak{D}_i(X,Y;\mathbb{G})$$

其中，使用梯度下降算法时，对应的影像合成模型的优化方向为

$$\frac{\partial \mathfrak{D}}{\partial \mathbb{G}} = \frac{\partial \mathfrak{D}_r}{\partial \mathbb{G}} + \frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}$$

当$\left\|\frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}\right\| > 0$时，优化会偏向任务无关的方向，合成的影像相对于任务的价值降低；特别当$\left\|\frac{\partial \mathfrak{D}_i}{\partial \mathbb{G}}\right\| \gg \left\|\frac{\partial \mathfrak{D}_r}{\partial \mathbb{G}}\right\|$时，优化完全朝向与任务无关的方向，合成的影像对该任务完全没有价值。在 Cohen 等人给出的例子中，GAN 合成的影像虽然看起来更加真实，但在一些问题下甚至不如只使用 MAE 损失有用[14]，原因就在于判别器的优化方向在任务无关的方向上的分量大于在任务相关方向上的分量。因此，如果需要将合成的影像用于某一下游任务，必须突出任务相关的分量。

为了增加任务相关方向的优化分量，最直接的做法就是使用具体的任务模型作为影像合成目标，即让合成影像具有和真实影像在一个任务相关模型上有相近的输出。设$\mathbb{F}: \mathfrak{M}_Y \to \mathfrak{M}_T$是面向某一任务(如分割、分类、分割，配准等)的模型，其输入为合成模型$\mathbb{G}$的目标模态$Y \in \mathcal{Y} \sim \mathfrak{M}_Y$，输出为$Y$对应的任务标签$T \in \mathcal{T} \sim \mathfrak{M}_T$，该模型的求解目标为$\mathfrak{D}_s(Y,T;\mathbb{F}): T = \mathbb{F}(Y)$，那么将$\mathfrak{D}_t(X,Y;\mathbb{G}) = \mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F})$加入到合成模型的优化目标中，形成扩展的的优化目标

$$\tilde{\mathfrak{D}}(X,Y;\mathbb{G}) = \mathfrak{D}(X,Y;\mathbb{G}) + \mathfrak{D}_t(X,Y;\mathbb{G}),$$

即可产生一个面向任务的偏置$\frac{\partial \mathfrak{D}_t}{\partial \mathbb{G}}$。实验表明，这样的偏置可以显著提高合成模型对任务的适用性，将合成的影像用于训练对下游任务也有一定的帮助。比如在风格转换任务中，使用多尺度结构相似性约束作为优化目标，可以显著改善图像超分辨和去噪后的失真(即提高了结构相似性)[16]；在从其他模态合成 CT 影像的任务中，使用阈值分割约束即可提高合成的 CT 影像中不同组织(骨头、软组织、脂肪、气体等)间的差异性[17]。

3.2 通过网络模型形成偏置

对于很多实际问题，往往需要使用一个复杂的模型来得到近似的解决方案。例如，用一个卷积网络来进行病灶分割、生存期预测等。尽管这些模型很可能无法达到足够好的性能，但只要它们产生的结果与真实结果之间的误差在可接受范围，依然可以用它们来为影像合成模型产生面向任务的偏置。

同时，除了上述$\mathfrak{D}_t(X,Y;\mathbb{G})=\mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F})$的偏置目标外，还可以通过替换其中的$T$得到偏置目标的另一种形式：$\mathfrak{D}_t(X,Y;\mathbb{G})=\mathfrak{D}_s\big(Y,\mathbb{F}(\mathbb{G}(X));\mathbb{F}\big)$。与前一种形式相比，这种形式不需要引入训练样本的监督信息，故而在求解合成模型时可以保持和原来相同的数据量。此外，如果任务模型过于复杂，还可以使用图 3-(4)的方式，通过提取任务模型中间层的特征，并使用特征一致性约束来减少计算量。在之前的工作中，我们使用的大都是这样一种形式[3,10,15]。需要注意的是，根据网络的特点，浅层的特征对任务的表达能力通常弱于深层的特征，因此使用浅层的特征形成的对任务的偏置通常也要比深层的特征弱一些。

3.3 嵌入任务模型中的影像合成

虽然使用与任务相关的偏置能够使合成的影像更加适合该任务，但并不总是容易得到一个好的任务模型。因此，我们也希望合成的影像能帮助提高任务模型的性能[15]。此时，可以通过优化模型

$$\min_{\mathbb{F}}\mathfrak{D}_s(\mathbb{G}(X),T;\mathbb{F}),X\in\mathcal{X}$$

来利用从$\mathcal{X}$合成的影像提高任务模型$\mathbb{F}$；同时，$\mathcal{Y}$中的影像也可以通过

$$\min_{\mathbb{F}}\mathfrak{D}_s(Y,T;\mathbb{F}),Y\in\mathcal{Y}$$

同时加入到对$\mathbb{F}$的优化中。此时，合成模型$\mathbb{G}$通过优化

$$\min_{\mathbb{G}}\widetilde{\mathfrak{D}}(X,Y;\mathbb{G})$$

得到，以使其产生对任务的偏置。在优化过程中，$\mathbb{F}$和$\mathbb{G}$的求解要联合进行，其中$\mathfrak{D}_t$项可以根据目标模态影像和标签的数量选择。这个模型为许多任务提供了新的解决思路。以跨模态配准为例，同模态(如T1-MRI($\mathfrak{M}_X$)和T1-MRI($\mathfrak{M}_T$))之间的配准已经有许多成熟的工具，但跨模态(如 FDG-PET($\mathfrak{M}_Y$)和 T1-MRI($\mathfrak{M}_T$))的配准($\mathbb{F}$)依然是一个极具挑战的课题。在这个问题上，我们可以根据 FDG-PET 合成 T1-MRI[18]，进而用同模态配准方法得到配准参数，再应用于 FDG-PET；也可以根据 T1-MRI 合成 FDG-PET($\mathbb{G}$)，进而联合求解$\mathbb{F}$和$\mathbb{G}$得到 FDG-PET 和 T1-MRI 间的模型。

另外，有的任务模型期望同时使用多种模态来达到更好的性能[15,19]，但部分训练样本只有其中一种模态的影像($\mathfrak{M}_X$)，即另一种模态的影像($\mathfrak{M}_Y$)缺失了。此时，我们可以使用合成模型($\mathbb{G}$)来生成缺失的影像，同时提高多模态任务模型(记为$\widetilde{\mathfrak{D}}_s(X,Y,T;\mathbb{F}):T=\mathbb{F}(X,Y)$)的性能。这种情况下，依然可以通过联合优化$\mathbb{F}$和$\mathbb{G}$来求解，只需要将此时的偏差目标改为

$$\mathfrak{D}_t(X,Y;\mathbb{G})=\widetilde{\mathfrak{D}}_s(X,\mathbb{G}(X),T;\mathbb{F})$$

以阿尔茨海默病的影像诊断为例，同时使用 MRI($\mathfrak{M}_X$)和 PET($\mathfrak{M}_Y$)被认为是达到更好性能的有效手段，但许多样本没有采集 PET 影像。因此，多模态 MRI-PET 诊断模型面临数据缺失的问题。使用上述方法便可补全缺失数据，利用所有的样本训练多模态诊断模型。

四、跨模态影像合成的应用

跨模态影像合成技术在成像、重建、配准、分割、预测、诊断等医学影像智能计算的各个任务中都有所应用，同时也涉及到 MRI、PET、CT、X-光平片等影像模态。当前，跨模态医学影像合成的应用场景包括但不限于影像转换、数据扩充、数据统一、隐私保护、可信智能等。从合成图像的目的来说，这些应用大致可分为影像替代和影像补全两类。

4.1 影像的等效替代

某些模态的影像由于条件限制难以获取，但又在实际应用中必不可少。这时，可以尝试利用其他可获得模态的影像合成出该模态的影像。例如，PET 成像通常依赖 CT 影像进行辐射衰减校正，但在 PET/MRI 设备中无法采集 CT 影像，此时可以利用 MRI 影像或者未校正的 PET 影像合成 CT 影像，用以完成 PET 校正。同时，有些模态的影像采集过程需要依赖耗费高或者耗时长的成像方式，难以在临床中广泛的使用。例如，脑血容量(CBV)是评价颅内占位性病变最有用的参数，但 CBV 的

测量依赖于血流灌注成像技术，存在成像时间长、成本高、给患者带来极大不适等明显缺点。考虑到 CBV 影像中的信息可能也部分的存在于其他 MRI 序列中，可以尝试利用多个 MRI 序列来合成 CBV 影像，从而获得一个近似的结果[20]。

4.2 缺失影像的补全

对于有些任务，使用多种模态的影像相互配合,可以达到更好的精度。例如，同时使用 MRI 和 PET 影像可以建立更加准确的阿尔茨海默病诊断模型[3,15]，同时使用多种 MRI 序列可以提升胶质瘤分割精度，以便显现更加完整的病灶面貌[21]。但由于病人意愿或者条件限制等原因，数据不完整的情况非常普遍。这时，可以通过影像合成技术使用已有模态的影像合成缺失模态的影像，从而利用所有的样本来训练模型，并可以将该模型应用于可能存在缺失模态影像的样本上。

需要注意的是，影像替代和影像补全虽然技术上是相似的，但在目的上有着本质的区别。前者假设源模态包含目标模态的完整信息，希望挖掘源模态与目标模态的共有信息，并将这种信息以目标模态的模式呈现出来。后者则假设不同模态中的信息是互补的，希望合成的目标模态影像在共有信息的基础上呈现出与已有模态影像互补的信息，只有这样合成影像才能在下游任务中发挥正向的作用。例如，在 CT 影像中，不同组织通常对应不同的 HU 值(辐射衰减系数)，因此在使用 MRI 影像合成 CT 影像时，希望 MRI 影像所反映的组织分布信息以 HU 值的形式呈现。而在脑肿瘤的分割中，不同模态的影像都只有肿瘤区域的一部分信息，合成的影像只有提供本模态特有的与其他模态互补的信息，才能提高分割精度。这是很难做到的，即便使用面向任务的合成技术也很难达到期望，多模态模型性能的提升通常是得益于样本数量的增加或者数据多样性的提高。

4.3 在数据统一中的应用

医学影像通常呈现出多样化的特点，即使同一种模态的影像，也可能存在分辨率、噪声、数值分布等方面较大的差异。因此，将训练数据集上获得的模型应用于跨中心数据时，其性能会出现难以预测的下降。如果将这种富含多样性的影像视为广义的多模态影像，就可以使用跨模态影像合成技术将多中心的数据变成统一的风格，从而提高任务模型跨中心的迁移能力[22]。

五、总结与展望

本文介绍了跨模态医学影像合成方法及其应用，首先简介了该项研究的意义和技术难点，然后回顾了相关技术的发展和应用场景。目前，跨模态医学影像合成已得到广泛的研究，并在医学影像智能计算的诸多方向中有所应用，给许多问题的解决带来了新思路。但是，当前技术合成的影像和真实影像仍然存在较大的差异，面临明显的失效风险。展望未来，将与应用相关的目标和合成技术相结合，从而产生面向特定应用的合成技术是一条值得探索的可行技术路线。其中，如何融入目标任务相关的知识、如何设计与目标任务相关的模型、如何对影像合成模型进行专业化的改进等，都值得进一步深入研究。

责任编委 储珺

参考文献

[1] AS Fard, DC Reutens, and V Vegh. CNNs and GANs in MRI-based cross-modality medical image estimation. ArXiv Preprint, 2021, abs/2106.02198.

[2] Y Pan, M Liu, C Lian, Y Xia, and D. Shen. Spatially-constrained Fisher representation for brain disease identification with incomplete multi-modal neuroimages. IEEE Transactions on Medical Imaging, 2020, 39(9):2965-2975.

[3] Y Pan, M Liu, Y Xia, and D Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[4] Y Huang, L Shao, and AF Frangi. Cross-modality image synthesis via weakly coupled and geometry co-regularized joint dictionary learning. IEEE Transactions on Medical Imaging, 2018, 37(3):815-827.

[5] A Jog, A Carass, S Roy, DL Pham, and JL Prince. Random forest regression for magnetic resonance image synthesis. Medical Image Analysis, 2017, 35:475–488.

[6] R Li, W Zhang, H Suk, L Wang, J Li, D Shen, and S Ji. Deep learning based imaging data completion for improved brain disease diagnosis. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2014, 17 Pt 3: 305-12.

[7] D Nie, X Cao, Y Gao, W Li, and D Shen. Estimating CT image from MRI data using 3D fully convolutional networks. Proceedings of MICCAI workshop on Deep Learning and Data Labeling for Medical Applications (DLMIA), Athens, Greece, October 21, 2016, 170-178.

[8] L Xiang, Y Qiao, D Nie, L An, W Lin, Q Wang, and D Shen. Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing, 2017, 267: 406-416.

[9] M Liu, T Breuel, and J Kautz. Unsupervised image-to-image translation networks. Advances in neural information processing systems, 2017.

[10] Y Pan and Y Xia. Ultimate Reconstruction: Understand your bones from orthogonal views. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), 2021, 1155-1158.

[11] F Isensee, PF Jaeger, SAA Kohl, J Petersen, and KH Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2021, 18:203–211.

[12] H Yang, P Qian, and C Fan. An indirect multimodal image registration and completion method guided by image synthesis. Computational and Mathematical Methods in Medicine, 2020.

[13] T Wang, M Liu, J Zhu, A Tao, J Kautz, and B Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, 8798-8807.

[14] JP Cohen, M Luck, and S Honari. Distribution matching losses can hallucinate features in medical image translation. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2018, 529–536.

[15] Y Pan, Y Chen, D Shen, and Y Xia. Collaborative image synthesis and disease diagnosis for classification of neurodegenerative disorders with incomplete multi-modal neuroimages. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2021, 480–489.

[16] R Malhotra, K Sharma, K Kumar, and N Rath. Integrating SSIM in GANs to generate high-quality brain MRI images. Data Engineering and Communication Technology. Springer, Singapore, 2021, 419-426.

[17] RR Colmeiro, C Verrastro, D Minsky, and T Grosges. Towards a whole body [18F] FDG positron emission tomography attenuation correction map synthesizing using deep neural networks. Journal of Computer Science and Technology, 2021.

[18] Q Yang, N Li, Z Zhao, X Fan, E Chang, and Y Xu. MRI cross-modality image-to-image translation. Scientific Reports. 2020, 10.

[19] J Wei, Y Pan, Y Xia, and D Shen. Learning to synthesize 7T MRI from 3T MRI with few data by deformable augmentation. MICCAI 2021 Workshop on Machine Learning in Medical Imaging (MLMI), 2021.

[20] Y Pan, J Huang, B Wang, P Zhao, Y Liu, and Y Xia. Cerebral blood volume prediction based on multi-modality magnetic resonance imaging. MICCAI 2021 Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), 2021.

[21] H Jia, Y Xia, W Cai, and H Huang. Learning High-Resolution and Efficient Non-local Features for Brain Glioma Segmentation in MR Images. Medical Image Computing and Computer Assisted Intervention (MICCAI), 2020, 480-490.

[22] H Lei, W Liu, H Xie, B Zhao, G Yue, and B. Lei, Unsupervised Domain Adaptation Based Image Synthesis and Feature Alignment for Joint Optic Disc and Cup Segmentation, IEEE Journal of Biomedical and Health Informatics, 2022 26(1): 90-102.

潘永生

上海科技大学生物医学工程学院博士。研究方向：图像合成、机器学习、疾病诊断。
Email: panysh@shanghaitech.edu.cn

崔恒飞

西北工业大学计算机学院副教授。研究方向：医学影像分析，模式识别。
Email: hfcui@nwpu.edu.cn

夏勇

西北工业大学计算机学院教授。研究方向：医学影像分析，图像处理，模式识别。
Email: yxia@nwpu.edu.cn

专题综述

用于物体位姿估计的端到端概率 PnP

同济大学 陈涵晟 田炜 熊璐

本文是同济大学团队解读其在CVPR 2022获得最佳学生论文奖的工作EPro-PnP[1]。论文研究的问题是基于单张图像估计物体在3D空间中的位姿。现有方法中，基于PnP几何优化的位姿估计方法往往通过深度网络提取2D-3D关联点，然而因为位姿最优解在反向传播时存在不可导的问题，难以实现以位姿误差作为损失对网络进行稳定的端到端训练，此时2D-3D关联点依赖其他代理损失的监督，这对于位姿估计而言不是最佳的训练目标。为解决这一问题，我们从理论出发，提出了EPro-PnP模块，其输出位姿的概率密度分布而非单一的位姿最优解，从而将不可导的最优位姿替换为了可导的概率密度，实现了稳定的端到端训练。EPro-PnP通用性强，适用于各类具体任务和数据，可以用于改进现有的基于PnP的位姿估计方法，也可以借助其灵活性训练全新的网络。从更一般的意义来说，EPro-PnP本质是将常见的分类softmax带入到了连续域，理论上可以推广至训练一般的嵌套了优化层的模型。

一、研究背景

我们研究的是 3D 视觉中的一个经典问题：基于单张 RGB 图像定位其中的 3D 物体。具体而言，给定一张含有 3D 物体投影的图像，我们的目标是确定物体坐标系到相机坐标系的刚体变换。这一刚体变换被称为物体的位姿，记作y，其包含两部分：1)位置(position)分量，可用 3x1 的位移向量t表示；2)朝向(orientation)分量，可用 3x3 的旋转矩阵R表示。

针对这一问题，现有方法可以分为显式和隐式两大类。显式方法也可称作直接位姿预测，即使用前馈网络(FFN)直接输出物体位姿的各个分量，通常是：1)预测物体的深度，2)找出物体中心点在图像上的2D投影位置，3)预测物体的朝向(朝向的具体处理方法可能比较复杂)。利用标有物体真实位姿的图像数据，可以设计损失函数直接监督位姿预测结果，轻松地实现网络的端到端训练，如图1所示。然而，这样的网络缺乏可解释性，在规模较小的数据集上易于过拟合。在3D目标检测任务中，显式方法占据主流[2, 3, 4]，尤其是对于规模较大的数据集，例如nuScenes[5]。

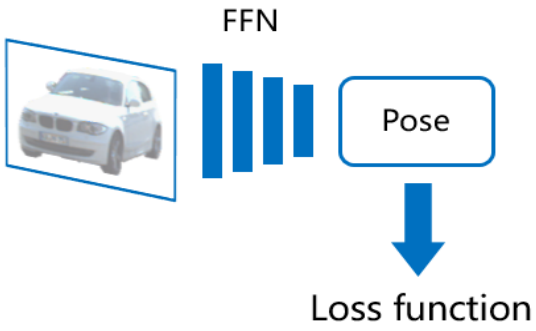


图 1 显示位姿估计网络结构示意图

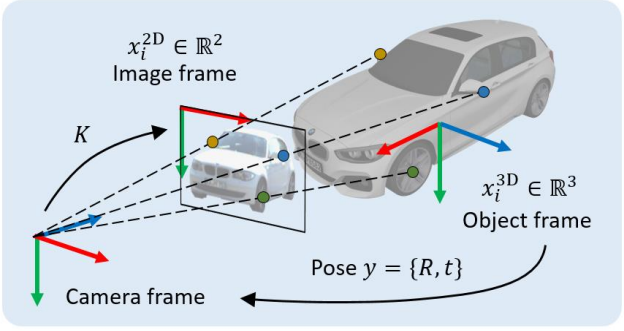


图 2 基于 PnP 的隐式位姿估计方法示意图

隐式方法则是基于几何优化的位姿估计方法，最典型的代表是基于 PnP 的位姿估计方法[6, 7, 8]。这类方法中，首先需要在图像坐标系中找出N个 2D 点(第i点 2D 坐标记作$x_i^{2D} \in \mathbb{R}^2$)，同时在物体坐标系中找出与之相关联的$N$个 3D 点(第$i$点 3D 坐标记作$x_i^{3D} \in \mathbb{R}^3$)，有时还需要获取各对点的关联权重(第$i$对点的关联权重记作$w_i^{2D} \in \mathbb{R}_+^2$)。如图 2 所示，根据透视投影约束，这$N$对 2D-3D 加权关联点隐式地定义了物体的最优位姿。具体而言，我们可以找出使重投影误差最小的物体位姿y^*：

$$y^* = \arg\min_y \frac{1}{2}\sum_i^N \|f_i(y)\|^2$$

其中$f_i(y) = w_i^{2D} \circ \left(\pi\left(Rx_i^{3D} + t\right) - x_i^{2D}\right)$，表示加权重投影误差，是位姿$y = \{R, t\}$的函数。$\pi(\cdot)$表示含有内参的相机投影函数，$\circ$表示元素乘积。PnP 方法常见于物体几何形状已知的 6 自由度位姿估计任务中。

如图 3 所示，基于 PnP 的方法也需要前馈网络去预测 2D-3D 关联点集$X := \{x_i^{3D}, x_i^{2D}, w_i^{2D} | i = 1 \cdots N\}$。相比于直接位姿预测，这一深度学习结合传统几何视觉算法的模型有非常好的可解释性，其泛化性能较为稳定，但在以往的工作中模型的训练方法存在缺陷。很多方法通过构建代理损失函数，去监督X这一中间结果，这对于位姿而言不是最优的目标。例如，已知物体形状的前提下，可以预先选取出物体的 3D 关键点，然后训练网络去找出对应的 2D 投影点位置[7]。这也意味着代理损失只能学习X中的部分变量，因此不够灵活。如果我们不知道训练集中物体的形状，需要从零开始学习X中的全部内容该怎么办？

显示和隐式方法的优势互补，如果能够通过监督 PnP 输出的位姿结果，端到端地训练网络去学习关联点集X，则可以将二者优势结合。为实现这一目标，一些近期研究利用隐函数求导实现了 PnP 层的反向传播[9, 10, 11]。然而，PnP 中的 argmin 函数在某些点是不连续不可导的，使得反向传播并不稳定，直接训练难以收敛。

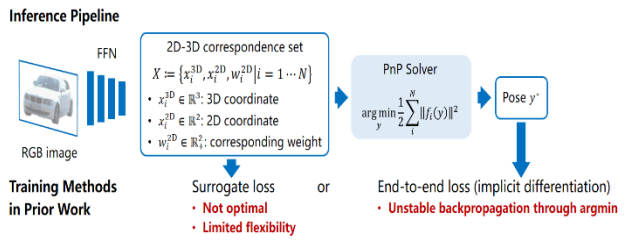


图 3　基于 PnP 的位姿估计网络结构及训练方法示意图

二、EPro-PnP方法介绍

1. EPro-PnP 模块

为了实现稳定的端到端训练，我们提出了端到端概率 PnP(end-to-end probabilistic perspective-n-point)，即 EPro-PnP，如图 4 所示。其基本思想是将隐式位姿视作一个概率分布，则其概率密度$p(y|X)$对于X是可导的。首先基于重投影误差定义位姿的似然函数：

$$p(X|y) = \exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2$$

若使用无信息先验，则位姿的后验概率密度为似然函数的归一化结果：

$$p(y|X) = \frac{p(X|y)}{\int p(X|y)\mathrm{d}y} = \frac{\exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2}{\int \exp -\frac{1}{2}\sum_i^N \|f_i(y)\|^2 \,\mathrm{d}y}$$

可以注意到，以上公式与常用的分类 softmax 公式($Softmax(a_i) = \exp a_i / \sum_j \exp a_j$)十分接近，其实 EPro-PnP 的本质就是将 softmax 从离散域搬到了连续域，把求和$\sum$换成了积分$\int$。

2. KL 散度损失

在训练模型的过程中，已知物体真实位姿y_{gt}，则可以定义目标位姿分布$t(y)$。此时可以计算KL散度$D_{\mathrm{KL}}(t(y)|p(y|X))$作为训练网络所用的损失函数(因$t(y)$固定，实际上也就是交叉熵损失函数)。在目标$t(y)$趋近于Dirac函数的情况下，基于KL散度的损失函数可

图 4　基于 EPro-PnP 的位姿估计网络结构示意图

以简化为以下形式：

$$L_{\mathrm{KL}} = \frac{1}{2}\sum_{i}^{N}\left\|f_i(y_{\mathrm{gt}})\right\|^2 + \log\int \exp -\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2\,\mathrm{d}y + const$$

如对其求导则有：

$$\frac{\partial L_{\mathrm{KL}}}{\partial(\cdot)} = \frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i}^{N}\left\|f_i(y_{\mathrm{gt}})\right\|^2 - \mathop{\mathbb{E}}_{y\sim p(y|X)}\frac{\partial}{\partial(\cdot)}\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2$$

可见，该损失函数由两项构成，第一项(记作L_{tgt})试图降低位姿真值y_{gt}的重投影误差，第二项(记作L_{pred})试图增大预测位姿$p(y|X)$各处的重投影误差。二者方向相反，效果如图 5(左)所示。作为类比，图 5(右)就是我们在训练分类网络是常用的分类交叉熵损失。

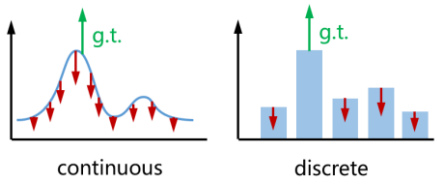


图 5　EPro-PnP 所用的连续损失与离散分类损失的类比

3. 蒙特卡洛位姿损失

需要注意到，KL 损失中的第二项$L_{\mathrm{pred}} = \log\int \exp -\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2\,dy$中含有积分，这一积分没有解析解，因此必须通过数值方法进行近似。综合考虑通用性，精确度和计算效率，我们采用蒙特卡洛方法，通过采样来模拟位姿分布。具体而言，我们采用了一种重要性采样算法——Adaptive Multiple Importance Sampling(AMIS)[12]，计算出K个带有权重v_j的位姿样本y_j，我们将这一过程称作蒙特卡洛 PnP：

$$\{y_j, v_j | j = 1\cdots K\} = PnP_{\mathrm{MC}}(X)$$

据此，第二项L_{pred}可以近似为关于权重v_j的函数，且v_j可以反向传播：

$$L_{\mathrm{pred}} = \log\int p(X|y)\mathrm{d}y \approx \log\frac{1}{K}\sum_{j=1}^{K} v_j$$

位姿采样的可视化效果如图 6 所示。

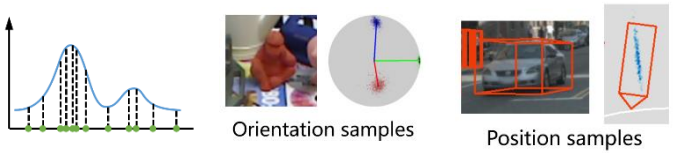


图 6　位姿采样示意图及可视化

4. 针对 PnP 求解器的导数正则化

尽管蒙特卡洛 PnP 损失可以用于训练网络得到高质量的位姿分布，但在推理阶段，还是需要通过 PnP 优化求解器来得到最优位姿解y^*。常用的高斯-牛顿及其衍生算法通过迭代优化求解y^*，其迭代增量是由代价函数$\frac{1}{2}\sum_{i=1}^{N}\|f_i(y)\|^2$的一阶和二阶导数决定的。为使 PnP 的解$y^*$更接近真值$y_{\mathrm{gt}}$，可以对代价函数的导数进行正则化。设计正则化损失函数如下：

$$L_{\mathrm{reg}} = l\left(y^* + \Delta y, y_{\mathrm{gt}}\right)$$

其中，Δy为高斯-牛顿迭代增量，与代价函数的一阶和二阶导数有关，且可以反向传播，$l(\cdot,\cdot)$表示距离度量，对于位置使用 smooth L1，对于朝向使用 cosine similarity。在y^*与y_{gt}不一致时，该损失函数促使迭代增量Δy指向实际真值。

三、基于EPro-PnP的位姿估计网络

我们在 6 自由度位姿估计和 3D 目标检测两个子任务上分别使用了不同的网络。其中，对于 6 自由度位姿估计，在 ICCV 2019 的 CDPN[8]基础上稍加修改并用 EPro-PnP 训练，用来进行消融实验；对于 3D 目标检测，在 ICCVW 2021 的 FCOS3D[13]基础上设计了全新的变形关联(deformable correspondence)检测头，以证明 EPro-PnP 可以训练网络在没有物体形状知识的情况下直接学出所有 2D-3D 点和关联权重，从而展现 EPro-PnP 在应用方面的灵活性。

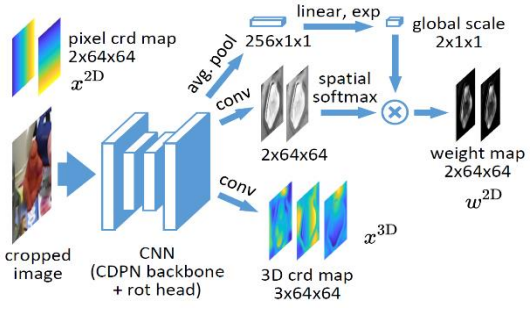


图 7　6 自由度位姿估计网络结构

1. 用于 6 自由度位姿估计的稠密关联网络

网络结构如图 7 所示，只是在原版 CDPN[8]的基础上修改了输出层。原版 CDPN 使用已经检测到的物体 2D 框裁剪出区域图像，输入到 ResNet34 backbone 中。原版 CDPN 将位置与朝向解耦为两个分支，位置分支使用直接预测的显式方法，而朝向分支使用稠密关联和 PnP 的隐式方法。为了研究 EPro-PnP，改动后的网络只保留了稠密关联分支，其输出为 3 通道的 3D 坐标图，以及 2 通道关联权重，其中关联权重经过了 spatial softmax 和 global weight scaling。增加 spatial softmax 目的是对权重w_i^{2D}进行归一化，使其具有类似 attention map 的性质，可以关注相对重要的区域，实验证明权重归一化也是稳定收敛的关键。Global weight scaling 反映了位姿分布$p(y|X)$的集中程度。该网络仅需 EPro-PnP 的蒙特卡洛位姿损失就可以训练，此外可以增加导数正则化，以及在物体形状已知的情况下增加额外的 3D 坐标回归损失。

2. 用于 3D 目标检测的变形关联网络

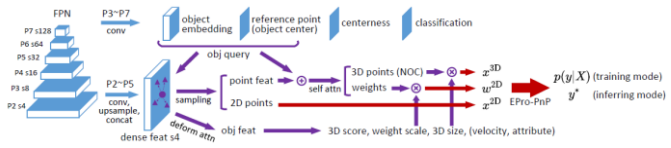


图 8　3D 目标检测网络结构

网络结构如图 8 所示。总体而言是基于 FCOS3D[13]检测器，参考 deformable DETR[14]设计的网络结构。在 FCOS3D 的基础上，保留其 centerless 和 classification 层，而将其原有的位姿预测层替换为 object embedding 和 reference point 层，用于生成 object query。参考 deformable DETR，我们通过预测相对于 reference point 的偏移量得到 2D 采样位置(也就得到了x_i^{2D})。采样后的 feature 经由 attention 操作聚合为 object feature，用于预测物体级别的结果(3D score，weight scale，3D box size 等)。此外，采样后各点的 feature 在加入 object embedding 并经由 self-attention 处理后输出各点所对应的的 3D 坐标x_i^{3D}和关联权重w_i^{2D}。所预测的$x_i^{3D}, x_i^{2D}, w_i^{2D}$全部可由 EPro-PnP 的蒙特卡洛位姿损失训练得到，不需要额外正则化就可以收敛并有较高的精度。在此基础上，可以增加导数正则化损失和辅助损失进一步提升精度(具体细节在我们论文的补充材料中给出)。

四、实验结果

1. 6 自由度位姿估计任务

表 1　6 自由度估计任务的实验结果

Method		ADD-0.1d
CDPN without translation head		74.54
+ Batch=32, LM solver (fair baseline)		79.96
Basic EPro-PnP Loss		**92.66 (+12.70)**
+ Tricks	+ Regularize derivatives	93.43
	+ Initialize from CDPN	95.76
	+ Long schedule (320 ep.)	95.80

使用 LineMOD[19]集进行实验，并严格与 CDPN baseline 进行比对，主要结果见表 1。可见，增加 EPro-PnP 损失进行端到端训练，精度显著提升(+12.70)。继续增加导数正则化损失，精度进一步提升。在此基础上，使用原版 CDPN 的训练结果初始化并增加 epoch(保持总 epoch 数与原版 CDPN 的完整三阶段训练一致)可以使精度进一步提升，其中预训练 CDPN 的优势部分来源于 CDPN 训练时有额外的 mask 监督。

表 2　与其它 6 自由度估计方法的比较

Method	Type	ADD-0.1d
CDPN	PnP + Explicit depth	89.86
HybridPose	Hybrid geometric constraints	91.3
GDRNet	PnP + Explicit depth	93.6
DPOD	PnP + Explicit refiner	95.15
EPro-PnP (ours)	PnP	95.80
PVNet-RePOSE	PnP + Implicit refiner	96.1

表 2 是 EPro-PnP 与各种领先方法[8, 15, 16, 17, 18]的比较。由较落后的 CDPN 改进而来的 EPro-PnP 在精度上接近 SOTA，并且 EPro-PnP 的架构简洁，完全基于 PnP 进行位姿估计，不需要额外进行显式深度估计或位姿精修，因此在效率上也有一定优势。

2. 3D 目标检测任务

使用 nuScenes[5]数据集进行实验，与其他方法对

表 3　3D 目标检测任务的实验结果

Method	Type	NDS↑	mAP↑	mATE↓	mAOE↓
MonoDIS	Explicit (direct prediction)	0.384	0.304	0.738	0.546
CenterNet	Explicit (direct prediction)	0.400	0.338	0.658	0.629
FCOS3D	Explicit (direct prediction)	0.428	0.358	0.690	0.452
PGD	Explicit + Ground constraint	0.448	**0.386**	0.626	0.451
EPro-PnP	PnP	**0.453**	0.373	**0.605**	**0.359**

比结果如表 3 所示。EPro-PnP 不仅相对 FCOS3D 有了明显提升，还超越了 FCOS3D 的另一个改进版本 PGD。更重要的是，EPro-PnP 目前是唯一在 nuScenes 数据集上使用几何优化方法估计位姿的。因 nuScenes 数据集规模较大，端到端训练的直接位姿估计网络已具有较好性能，而我们的结果说明了端到端地训练基于几何优化的模型能做到在大数据集上取得更加优异的性能。

3. 可视化分析

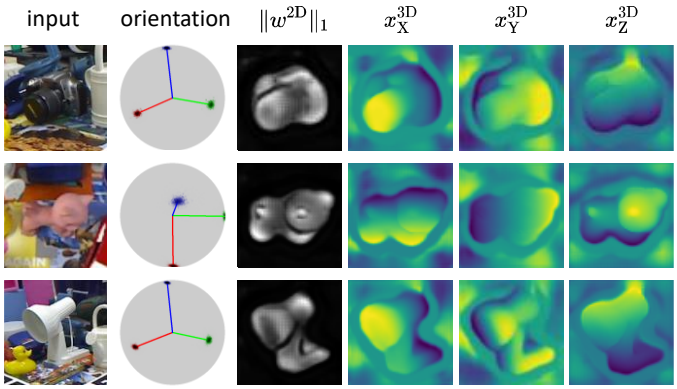


图 9　稠密关联网络的预测结果

图 9 显示了用 EPro-PnP 训练的稠密关联网络的预测结果。其中，关联权重图$\|w^{2D}\|_1$对图像中的重要区域进行了高光，类似于 attention 机制。由损失函数分析可知，高光区域对应的是重投影不确定性较低以及对位姿变动较为敏感的区域。

3D 目标检测的结果如图 10 所示。其中左上视图显示了变形关联网络采样出的 2D 点位置，红色表示w_i^{2D}水平 X 分量较高的点，绿色表示w_i^{2D}垂直 Y 分量较高的点。绿色点一般位于物体上下两端，其主要作用是通过物体高度来推算物体的距离，这一特性并非人为指定，完全是自由训练的结果。右图显示了俯视图上的检测结果，其中蓝色云图表示物体中心点位置的分布密度，反映了物体定位的不确定性。一般远处的物体定位不确定

图 10　3D 目标检测结果

性大于近处的物体。

EPro-PnP 的另一重要优势在于，能够通过预测复杂的多峰分布来表示朝向的模糊性。如图 11 所示，Barrier 由于物体本身旋转对称，朝向经常出现相差 180° 的两个峰值；Cone 本身没有特定的朝向，因此预测结果在各个方向均有分布；Pedestrian 虽不完全旋转对称，但因图像不清晰，不易判断正面和背面，有时也会出现两个峰值。这一概率特性使得 EPro-PnP 对于对称物体不需要在损失函数上做任何特殊处理。

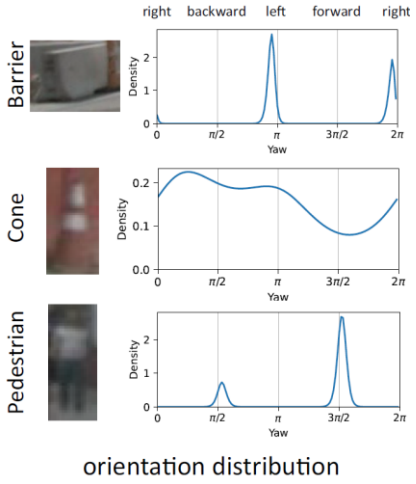


图 11　3D 目标检测网络预测的模糊朝向

四、总结

EPro-PnP 将原本不可导的最优位姿转变为可导的位姿概率密度，使得基于 PnP 几何优化的位姿估计网络可实现稳定且灵活的端到端训练。EPro-PnP 可应用于一般的 3D 物体位姿估计问题，即使在未知 3D 物体几何形状的情况下，也可以通过端到端训练学习得到物体

的 2D-3D 关联点。因此，EPro-PnP 拓宽了网络设计的可能性，例如我们提出的变形关联网络，这在以往是不可能训练的。此外，EPro-PnP 也可以直接被用于改进现有的基于 PnP 的位姿估计方法，通过端到端训练释放现有网络的潜力，提升位姿估计精度。从更一般的意义来说，EPro-PnP 本质是将常见的分类 softmax 带入到了连续域，不仅可用于其它基于几何优化的 3D 视觉问题，理论上还可以推广至训练一般的嵌套优化层的模型。

责任编委 王金甲

参考文献

[1] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, Hao Li. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. In CVPR 2022.

[2] Ze Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, Adrien Gaidon. Is Pseudo-Lidar Needed for Monocular 3D Object Detection? In ICCV, 2021.

[3] Tai Wang, Xinge Zhu, Jiangmiao Pang, Dahua Lin. Probabilistic and Geometric Depth: Detecting Objects in Perspective. In Conference on Robot Learning (CoRL), 2021.

[4] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, Justin Solomon. DETR3D: 3D Object Detection from Multi-View Images via 3D-to-2D Queries. In Conference on Robot Learning (CoRL), 2021.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In CVPR, 2020.

[6] Mahdi Rad, Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects without Using Depth. In ICCV, 2017.

[7] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, Hujun Bao. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In CVPR, 2019.

[8] Zhigang Li, Gu Wang, Xiangyang Ji. CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation. In ICCV, 2019.

[9] Dylan Campbell, Liu Liu, Stephen Gould. Solving the Blind Perspective-n-Point Problem End-to-End with Robust Differentiable Geometric Optimization. In ECCV, 2020.

[10] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, Tat-Jun Chin. End-to-End Learnable Geometric Vision by Backpropagating PnP Optimization. In CVPR, 2020.

[11] Eric Brachmann, Carsten Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In CVPR, 2018.

[12] Jean-Marie Cornuet, Jean-Michel Marin, Antonietta Mira, Christian P. Robert. Adaptive Multiple Importance Sampling. Scandinavian Journal of Statistics, 39(4):798–812, 2012.

[13] Tai Wang, Xinge Zhu, Jiangmiao Pang, Dahua Lin. FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. In ICCV Workshops, 2021.

[14] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In ICLR, 2021.

[15] Chen Song, Jiaru Song, Qixing Huang. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In CVPR, 2020.

[16] Gu Wang, Fabian Manhardt, Federico Tombari, Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In CVPR, 2021.

[17] Sergey Zakharov, Ivan Shugurov, Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In ICCV, 2019.

[18] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, Kris M. Kitani. RePOSE: Fast 6D Object Pose Refinement via Deep Texture Rendering. In ICCV, 2021.

[19] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, Vincent Lepetit. Multimodal Templates for Real-Time Detection of Texture-less Objects in Heavily Cluttered Scenes. In ICCV, 2011.

陈涵晟

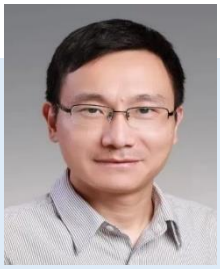
同济大学汽车学院 2020 级硕士研究生，导师为熊璐教授，副导师为田炜助理教授，主要研究方向为 3D 计算机视觉。

Email:hanshengchen97@gmail.com

田炜

博士毕业于德国卡尔斯鲁厄理工学院，现任同济大学汽车学院助理教授、硕士生导师，主要研究方向为面向智能驾驶的环境目标感知技术和轨迹预测技术，上海市浦江人才计划入选者，主持国家自然科学基金青年项目、上海市自然科学基金面上项目等横纵向课题，曾参与德国联邦教育研究部、德国博世研究院智能驾驶项目开发，并担任国际会议 IEEE ITSC2015、FUSION2021、CVCI2021 分会场主席，发表智能驾驶领域 SCI/EI 论文近 40 篇，著有专著 2 部。

Email: tian_wei@tongji.edu.cn



熊璐

工学博士、教授、博士生导师。现任同济大学新能源汽车工程中心副主任。长期从事汽车底盘控制、分布式驱动电动汽车动力学控制、智能驾驶相关科研工作，主持和参与国家重点研发计划项目、国家自然科学基金项目、973 计划、863 计划和国家支撑计划等多项国家和省部级项目；发表 SCI/EI 论文 100 余篇，授权专利 40 余项，参撰英文著作 2 部；曾获 2011 年中国汽车工业科技进步三等奖、2013 年上海市科技进步一等奖、2019 年上海市科技进步一等奖等多项奖励。任《同济大学学报》编委和国内外多个期刊的评审专家、国家自然基金和科技部重点研发计划等项目评审专家，担任国际汽车工程师学会(SAE)智能网联汽车技术委员会联合主席、中国汽车工程学会汽车智能交通分会副秘书长、中国汽车工程学会青年委员会副主任委员、中国自动化学会车辆控制与智能化专委会委员。

Email: xiong_lu@tongji.edu.cn

热点追踪

DINE：基于黑盒模型的无监督领域自适应学习

中科院自动化研究所 梁坚 赫然 新加坡国立大学 胡大鹏 冯佳时

一、摘要

为了减轻对标注数据的依赖，无监督领域自适应学习旨在将已有相关标记数据集(源域)中的知识转移到新的未标记数据集(目标域)上。现有的方法需要访问原始的源域数据，并依赖于其中的信息以识别目标样本，这一设置在数据隐私愈发重要的场景下难以得到有效的部署。近两年来，有少量研究希望利用在源数据上学习得到的白盒源模型代替源域数据来进行目标域数据的自适应学习，但这一方式依旧存在模型遭受逆生成攻击而泄露数据的风险。本文探讨了一种有趣的无监督领域自适应的问题设置，即在目标域自适应期间只能接触黑盒源模型(即只有网络预测可见)。具体地，我们提出了一种称为 DINE 的两步知识自适应学习框架。相关成果被 CVPR 2022 录用为口头报告。

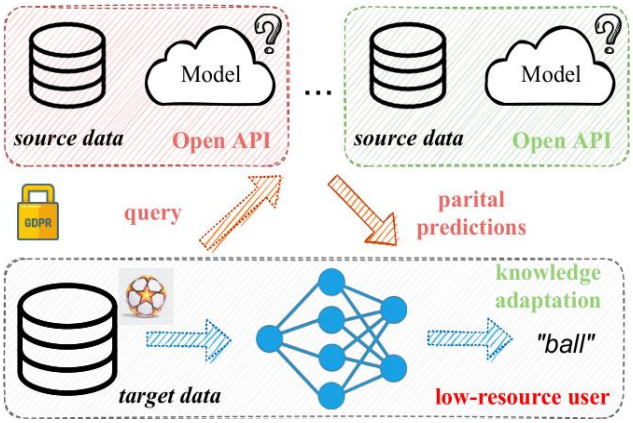


图 1 基于黑盒源模型的无监督领域自适应问题

二、引言

无监督领域自适应学习旨在将已有相关标记数据集(源域)中的知识转移到新的未标记数据集(目标域)上。现有领域自适应方法需要访问原始的源数据，它们通常使用领域对抗性训练[1]或最大平均差异最小化[2]等手段来对齐源域与目标域的特征分布。然而传统领域自

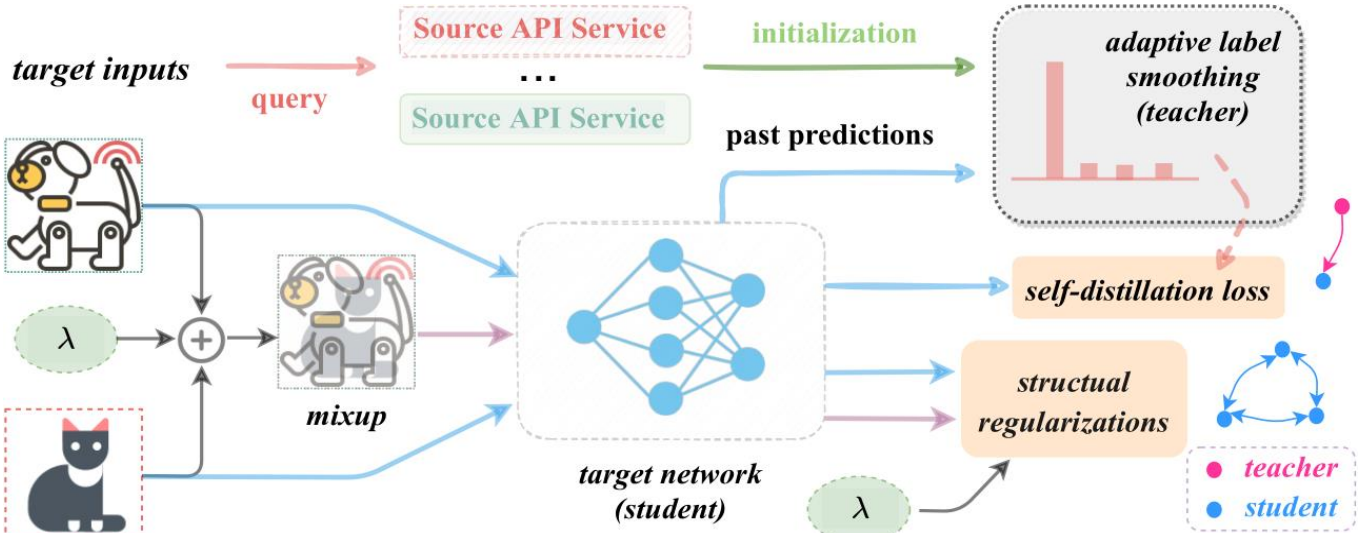


图 2 基于结构化知识蒸馏的无监督领域自适应方法整体框架

适应方法不能很好地应用于对数据安全比较重视的场景(如个人医疗信息、网络浏览历史等隐私数据)。近些年，有研究[3]提出使用在源数据上训练好的模型而非源数据来进行无监督目标域的领域自适应学习，取得了媲美数据依赖方法的识别性能。然而这种基于白盒模型进行知识迁移的方式依旧存在着被对抗生成学习等技术攻击导致数据泄露的风险。为了更好保护源域数据的数据安全，本文研究了基于黑盒模型的领域自适应学习问题(如图 1 所示)。在学习过程中，源数据和源模型信息均无法访问，用户仅能利用源模型的输出概率分布来进行知识迁移。同时，我们受知识蒸馏[4]框架的启发，提出了一种结构化知识蒸馏的新颖方法，在有效去除含噪教师信息的同时充分考虑了目前域数据的潜在结构化约束。此外，新方案不再需要目标域网络架构同源域架构一致，可以在资源受限的客户端采用轻量级网络架构进行知识抽取与整合。大量实验结果表明，我们提出的 DINE 框架在仅使用黑盒模型的条件下可以取得媲美依赖于源数据及基于白盒模型的迁移方法识别性能。

三、正文

本文所提出的 DINE 框架主要由两部分构成，即知识蒸馏阶段以及模型微调阶段，其中第一步为结构化知识蒸馏阶段，整体方案如图 2 所示。

在蒸馏阶段，我们采用适应性自知识蒸馏方法让目标(学生)模型来学习源(教师)模型的输出预测。我们将目标实例在多个源模型上的输出概率分布的平均作为指导分布，最小化目标模型与指导分布之间的 KL 散度损失。然而，由于目标域与源域的差异性，源模型的输出概率并不完全可信，因此我们提出一种自适应的标签平滑策略以调整来自源模型的输出分布。具体来说，我们保留源模型输出分布当中最大的 r 个值(r 默认为 1)，并且将其余类的概率修改为同一个值。通过这种方式，模型能够更加关注到最大的值并忽略部分噪声。不同于伪标签策略，我们不完全依赖于有噪声的伪标签，而是利用最大值作为置信度。为了进一步消除源模型预测中的噪声，我们用源模型的输出分布均值与目标模型预测分布之间的指数移动平均值作为指导分布，形成最终的自蒸馏损失。

此外，为了利用目标域中数据的结构化信息，我们对蒸馏过程进行结构正则化来约束知识蒸馏的过程。首先，我们通过 MixUp[5]来利用成对结构信息，通过最小化成对样本的混合输入在目标模型上的输出分布与一对样本输出分布的混合之间的交叉熵损失来优化网络。在此基础上，我们还考虑了蒸馏过程中目标域的全局结构信息。在蒸馏过程中，由于类别不平衡问题使得部分类相对容易学习，这可能会导致目标模型错误地将一些混淆的目标样本识别为此类。因此我们最大化模型输入与输出之间的互信息以鼓励样本总体的标签分布均匀。最后，在微调阶段，我们再一次应用互信息最大化来进一步精炼上一步蒸馏后的目标模型。

四、实验结果

表 1 DINE 及其他算法在 Office 数据库上的识别结果

Method	Type	A→D	A→W	D→A	D→W	W→A	W→D	Avg.
No Adapt.	Pred.	79.9	76.6	56.4	92.8	60.9	98.5	77.5
NLL-OT	Pred.	88.8	85.5	64.6	95.1	66.7	98.7	83.2
NLL-KL	Pred.	89.4	86.8	65.1	94.8	67.1	98.7	83.6
HD-SHOT	Pred.	86.5	83.1	66.1	95.1	68.9	98.1	83.0
SD-SHOT	Pred.	89.2	83.7	67.9	95.3	71.1	97.1	84.1
DINE	Pred.	91.6	86.8	72.2	96.2	73.3	98.6	86.4
DINE (full)	Pred.	91.7	87.5	72.9	96.3	73.7	98.5	86.7
ResNet-50↑↑, ViT↓↓ (source backbone) → **ResNet-50** (target backbone)								
No Adapt.	Pred.	88.2	89.2	74.5	97.2	77.2	99.3	87.6
NLL-OT	Pred.	91.3	91.4	76.4	97.2	78.2	99.4	89.0
NLL-KL	Pred.	91.7	91.8	76.3	97.2	78.4	99.0	89.1
HD-SHOT	Pred.	88.8	90.9	75.3	97.7	77.7	99.5	88.3
SD-SHOT	Pred.	91.6	92.8	77.8	98.7	78.5	**99.7**	89.8
DINE	Pred.	94.2	94.6	80.7	**98.8**	81.5	99.5	91.6
DINE (full)	Pred.	**95.5**	**94.8**	**81.2**	98.5	**82.0**	**99.7**	**91.9**

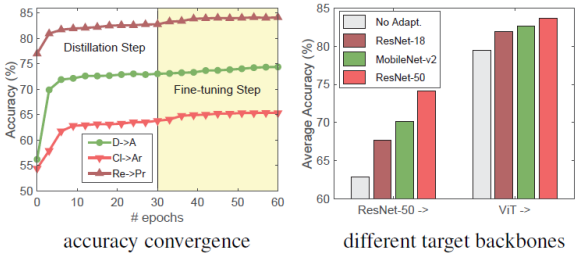


图 3 DINE 的收敛性分析及其对网络架构的敏感度分析

表 1 展示了我们的 DINE 和一些基准方法在 Office 数据库的识别准确率，其中上、下两部分分别为使用 ResNet50 和 ViT 作为源域基础架构进行迁移的结果。可以直观地发现，DINE 具有很大的性能优势。图 3 进一步展示了 DINE 在两个阶段的收敛性以及整体算法对目标域不同架构的敏感度。

责任编委 崔海楠

参考文献

[1] Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-adversarial training of neural networks." Journal of Machine Learning Research 17, no. 1 (2016): 2096-2030.

[2] Long, Mingsheng, Yue Cao, Jianmin Wang, and Michael Jordan. "Learning transferable features with deep adaptation networks." In International Conference on Machine Learning, pp. 97-105. PMLR, 2015.

[3] Liang, Jian, Dapeng Hu, and Jiashi Feng. "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation." In International Conference on Machine Learning, pp. 6028-6039. PMLR, 2020.

[4] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 2 (2015)..

[5] Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. "mixup: Beyond empirical risk minimization." arXiv preprint arXiv:1710.09412 (2017).

梁坚

中科院自动化研究所副研究员。主要研究方向为领域自适应、特征表示及迁移学习等。
Email: jian.liang@nlpr.ia.ac.cn

赫然

中科院自动化研究所研究员。主要研究方向为视觉内容生成、生物特征识别、机器学习等。
Email: rhe@nlpr.ia.ac.cn

顶会观察

CVPR 2022

南京理工大学　李俊

国际计算机视觉与模式识别大会(IEEE/CVF Computer Vision and Pattern Recognition Conference，CVPR)是计算机视觉和模式识别的顶级会议之一，与 ICCV 和 ECCV 并称为计算机视觉领域三大顶会。CVPR 的学术影响力越来越大，不仅是中国计算机学会推荐的人工智能领域 A 类国际学术会议，而且还在 Google Scholar 期刊与会议影响力榜单中排名第四(仅次于 Nature、NEJM 和 Science)。相比于去年，今年大会组委会仍有不少华人学者：香港科技大学权龙教授担任 General Chair，Wormpex AI Research 副总裁华刚博士担任 Program Chair，Google 研究科学家 Boqing Gong 担任 Tutorials Chair，旷视科技首席科学家孙剑博士等百名华人担任 Area Chairs。本届 CVPR 大会于 2022 年 6 月 19 日至 24 日在美国路易斯安那州新奥尔良举办，包括 4 天的正会和 2 天的 Workshops & Tutorials。

一、 会议亮点

线下线上混合形式：自新冠疫情流行以来，CVPR 2022 首次线下举办，无法线下参会人员仍可选择线上参会。据主办方统计，截至大会开幕，共有 9981 人注册参会，其中 5641 人现场参会，4340 人以线上方式参会。今年与以往不同，每篇论文都要以海报形式和虚拟会议两种方式进行工作展示及技术交流。海报形式：对于线下参会的作者，与 2019 年以前一样海报张贴；对于线上参会的作者，主办方提供了海报打印服务并由志愿者张贴。虚拟会议：作者准备一个五分钟的预录制视频和一张海报的文件，在会议平台上展示工作。为了方便交流，主办方还为每篇论文安排了线上、线下、同步和异步方式与参会者进行充分的交流。

严格的评审机制：今年会议组织方邀请了 304 位专家作为领域主席(area chair)，包括 4 位紧急领域主席。同时，组织方还邀请了 6427 位从业者作为审稿人参与论文评审，包括 1723 位紧急审稿人。在经验分布上，学生审稿人占比 31%，教职/研究人员审稿人占比为 69%。更重要的是，每篇论文会由 3 位 AC 一起处理(包括至少一位资深 AC 和至少两位领域内专家)，其中 AC 的分配是由 TPMS 和学科领域自动匹配产生。每篇论文至少分配 3 位审稿人，其中至少 2 位为主 AC 推荐。主 AC 负责主持审稿线上会议，AC 之间相互审核报告，核查错误，并详细讨论审稿意见，最终向作者发出通知。

开源代码：随着 CVPR 影响力的进一步提升，越来越多的研究工作发布了代码或数据，开源已经成为趋势。据 GitHub 项目的不完全统计，CVPR 2020 和 CVPR 2021 均有超 300 篇论文开源了代码。更惊喜的是，据 paperdigest 团队统计，今年已经超过 600 篇论文公开了源代码。代码开源的优势明显，免费透明，不仅可以增加研究者之间的协作机会，而且还能提升研究工作的影响力。当然，也存在一定的未知风险，这需要大家共同努力完善开源代码。

除了上述会议亮点外，大会现场缅怀了孙剑博士，通过一段视频带领大家一起回忆了孙剑博士科研成长之路。最佳(学生)论文、Longuet-Higgins 奖，青年研究者奖，Thomas S. Huang 纪念奖将在后面介绍。

二、 录用情况

CVPR 2022 收到的有效投稿和录用数量都有显著提高，大会共收到了 8161 篇有效投稿，最终接收了 2064 篇论文，接收率约为 25.3%。相较于 2021 年，今年 CVPR 的投稿量提升 15%(1068 篇)，录用率也有所上升，录用论文数量提升约为 24.3%(403 篇)。其中，有 342 篇论文录用为 Oral Presentations，比去年增加 47 篇，Oral 率约为 4.1%(与去年基本持平)。CVPR2022 会议涵盖的方向包括：识别(检测、分类与检索)、3D 视觉、图像与视频的生成、底层视觉(超分、恢复、去雾等)、深度学习与表示学习、视频分析与理解、视觉与语言、迁移学习、计算摄影、姿态估计与跟踪、场景分析和理解、无监督学习、行为识别、数据集与评估等方向。在 CVPR 2022 录用的论文中，识别、3D 视觉、图像与视频的生成、深度学习与表示学习四个方向的论文数量最多，均超过 150 篇。值得关注的是，来自中国学术机构与工业界企业取得了相当不俗的业绩。据公开接受列表，按单位统计，清华大学共 113 篇论文入选，中国科学院共有 91 篇论文入选，腾讯有 83 篇论文入选，阿里巴巴也有 67 篇论文入选。另据报道，商汤科技及联合实验室共 71 篇论文入选、南洋理工大学与香港理工大学多媒体实验室吕健勤教授团队有 18 篇论文入选、上海交通大学马利庄教授团队有 14 篇论文入选等。此外，谷歌与脸书在本次会议中仍有不错表现，分别有 89 篇和 73 篇论文入选。

三、 主题报告

本次 CVPR 2022 会议邀请了三位 Keynote 演讲者，报告内容涵盖类人智能、整合 AI 和视觉外观理解。

Learning to See the Human Way. 麻省理工学院脑与认知科学系教授 Josh Tennebaum 报告并讨论了如何建立类人智能的计算机视觉系统。近年来，计算机视觉是已经取得了注目的成绩，然而我们仍然无法拥有类人智能的机器系统(观察一张图或真实世界，就能轻松认识一切事物)。报告者受人类视觉和视觉认知系统的启发，引导出一种交替方法去建构可实践的机器视觉系统。基于可微与概率编程技术，此方法利用一些经典范式(如逆图形、综合分析、最佳解释推断等)理解视觉。虽然报告者展示了少许机器视觉成功的例子，但不幸的是，建构像人一样观察世界的机器系统，依旧困难重重。最后，报告者还指出了这一领域一些大挑战任务。

Toward Integrative AI with Computer Vision. Microsoft 技术会士、Azure AI 首席技术官黄学东描述并探讨了面向计算机视觉的整合 AI。当不相关的 AI 任务数量快速增长时，如何整合 AI 成为了一个重要的问题。报告者分享了一种多语言多模式的整合 AI 方法，即利用完整的语义表示统一在语音、语言和视觉的多种任务。特别是，当应用到计算机视觉任务时，报告者正在发展一种基础模型(Florence)，通过大尺度的图像和语言预训练，提出语义层概念，并在共同视觉任务(如识别、检测、分割等)中提升零样本/少样本的学习能力。通过搭建语义表示与视觉下游任务的关系，Florence 不仅在 COCO, VQA and Kinetcs-600 等数据库上取得了 state-of-the-art 实验结果，而且还发现了图形理解的新结果。最后，报告者还预见语义层将赋予计算机视觉有超越视觉感知的能力，而且能够流畅地连接像素到人类智能的核心(意图、推理、决策)。

Understanding Visual Appearance from Micron to Global Scale. 康奈尔大学 Ann S. Bowers 计算与信息科学学院院长 Kavita Bala 介绍了从微观到宏观的视觉外观理解。计算机视觉的核心是，利用视觉输入，让计算机理解和发掘我们生活的真实世界，包括形状、材质、场景、活动等。同时，我们可以在微观或宏观尺度下观察这个真实世界，例如：微米级的 CT 图、行星级的卫星图。因此，我们希望计算机既能从微观角度深层理解个体目标外观，又可以宏观地理解世界规模的事件。在本报告中，Kavita 介绍了团队在视觉理解方向的最新成果，包括真实视觉外观与渲染的图模型，形状与材质的重建，世界规模发现的视觉模式等。最后，报告者还探讨了这一领域的发展趋势。

四、 热点论文

2022 年度最佳论文奖评审委员会由 10 名国际权威学者组成，其中包括两名华人学者：北京大学的林宙辰教授和宾夕法尼亚州立大学 Yanxi Liu 教授。相比 2021 年，增加了两位评审委员。本年度大会共评选出

了 1 篇最佳论文，1 篇最佳学生论文，1 篇最佳论文提名，1 篇最佳学生论文提名。

最佳论文：Learning to Solve Hard Minimal Problems[1]，来自苏黎世联邦理工学院、华盛顿大学、佐治亚理工学院和捷克理工大学。在计算机视觉中，很多任务都面临困难的几何优化问题，例如 3D 重建、图像匹配、视觉里程计等。这类问题通常会简化为含有许多无意义解的最小化问题。本文提出了一种学习策略去选择一个起始的(问题-解)对，可以有效地找到问题的兴趣解，进而避免大量无意义解的计算。本文还开发了一个 RANSAC 求解器来计算三个校准相机的相对位置，进而证实了此学习策略的有效性。

最佳论文提名：Dual-Shutter Optical Vibration Sensing[2]，来自卡内基梅隆大学。视觉振动测量技术在声学、材料学科和生命科学领域中有着广泛的应用，如：远程捕捉音频、材质的物理属性、人体心率等。基于工作频率为 130Hz 的传感器，本文提出了一种新的视觉振动测量方法，不仅可以感知高速(高达 63kHz)的振动，而且还能同时感知多个场景源。此方法依赖于同时使用两个分别装有滚动和全局快门传感器的相机来捕捉场景。其中，滚动快门相机可以捕捉反映高速物体振动的扭曲斑点图像；全局快门相机能捕捉斑点图案的未扭曲参考图像。最后，通过捕捉由音源(如扬声器、人声和乐器)产生的振动，验证了这种新的测量方法。

最佳学生论文：EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation[3]，来自同济大学和阿里巴巴集团。从单图像中定位 3D 物体是计算机视觉中长期存在且非常重要的问题。本文提出了一种新颖的 Perspective-n-Points(PnP)方法，即用于一般端到端姿势估计的概率 PnP 层，它可以输出 SE(3)流形上的姿势分布，本质上是将分类 Softmax 引入连续域。2D-3D 坐标和对应的权重被视为中间变量，并通过最小化预测和目标姿势分布之间的 KL 散度来学习。此方法的效果明显优于有竞争力的基准方法，并且在 LineMOD 6DoF 姿势估计和 nuScenes 三维物体检测基准上，缩小了基于 PnP 的方法与特定任务最佳性能之间的差距。

最佳学生论文提名：Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields[4]，来自哈佛大学和谷歌研究院。近年来，神经辐射场(NeRF)是一种流行的视图合成技术。虽然它擅长表现具有平滑变化且视线依赖性外观的精细几何结构，但往往不能准确捕捉和重现光泽表面的外观。本文提出了一种反射 NeRF(Ref-NeRF)，通过引入反射辐射的表示方法代替了 NeRF 的 MLP 网络参数，并使用一系列空间变化的场景属性来构造这一网络函数。结果表明，结合法线向量的正则器，Ref-NeRF 显著提升了镜面反射的真实性和准确性。

另外，还有 29 篇论文入选最佳论文入围名单，其中华人学者为第一作者的论文数量超过半数，并且多篇论文也引起了广泛的讨论。脸书 AI 研究院的论文 Masked Autoencoders Are Scalable Vision Learners[5] 利用随机遮盖输入图片的子块，重建丢失块像素，由此来预训练模型，使得模型在下游任务中具有很好的泛化性能。中国科学院大学的 AnyFace: Free-style Text-to-Face Synthesis and Manipulation[6] 提出了一种风格自由的文本到人脸方法，去合成并编辑人脸。香港城市大学、德国马普所和斯坦福大学的 Learning to Deblur Using Light Field Generated and Real Defocus Images[7] 提出了一种新颖的去焦距去模糊网络，它可以利用强度并克服光照的缺点。

五、 大会获奖和竞赛奖

Longuet-Higgins Prize。该奖以理论化学家和认知科学家 H. Christopher Longuet-Higgins 的名字命名，它是由 PAMI 技术委员会颁发的计算机视觉基础贡献奖，表彰十年前对计算机视觉研究产生了重大影响的 CVPR 论文。今年获奖论文是来自卡尔斯鲁厄理工学院和丰田工业大学芝加哥分校的论文 Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite。

Young Researcher Awards。该奖旨在表彰计算机视觉领域的青年研究人员，鼓励年轻科学家继续做出开创性工作。今年获奖者分别是康奈尔大学计算机科学系助理教授 Bharath Hariharan 和普林斯顿大学计算机

科学系助理教授 Olga Russakovsky。

Thomas S. Huang 纪念奖。 自 2021 年起，为了缅怀一代宗师、华人计算机视觉泰斗 Thomas S. Huang(黄煦涛)教授，由 PAMITC 奖励委员会选出，以表彰在计算机视觉研究、教育和服务方面被公认为楷模的研究人员。今年第二届获奖者是斯坦福大学计算机科学系教授李飞飞。研究领域包括认识启发的 AI，机器学习、深度学习、计算机视觉和 AI+医疗健康。

竞赛奖。 在 CVPR 2022 研讨会上举办的各项挑战赛中，国内学术界和工业界都取得了不俗的成绩。西安电子科技大学在Woodscape鱼眼目标检测挑战和农业视觉 CropHarvest 竞赛中，分别获得了冠军。南京理工大学在 NTIRE 的竞赛中，获得了高效超分辨率赛道和 4 倍超分辨率赛道的双冠军。腾讯太极团队在轻量化 NSA 竞赛中，获得了超网络赛道冠军。美团在届细粒度视觉分类竞赛中，获得了植物标本识别赛道和大规模跨模态商品图像召回赛道的双冠军。联想研究院在 BDD100K MOT 挑战赛中，获得平均多目标跟踪准确度的冠军。

六、 总结展望

本年度 CVPR 大会中识别、3D 视觉、图像与视频的生成、深度学习、表示学习、Transformer 等领域依旧保持高热度。相比于 2021 年，底层视觉热度显著回升，计算摄影热度有所下降。更值得关注的是，计算机通过视觉传感器感知并认知真实物理世界，CVPR 越来越注重解决真实场景下的视觉问题，从 2D 到 3D，从微观到宏观，从识别到类人智能，甚至是超越类人智能。从第三方角度观察，计算机视觉领域还面临许多挑战：从物理世界到传感器与机器认知过程中，现有方法的能力上界是什么、人类已经发现的规律如何辅助机器认知物理世界、机器如何探知人类未曾发现的物理世界规律等。笔者认为回答好上述问题将会是计算机视觉的新机遇，从而更好地迈向更高层级的智能。

责任编委 魏秀参

参考文献

[1] Petr Hruby, Timothy Duff, Anton Leykin, Tomas Pajdla. Learning to Solve Hard Minimal Problems. CVPR 2022.

[2] Mark Sheinin, Dorian Chan, Matthew O'Toole, and Srinivasa G. Narasimhan. Dual-Shutter Optical Vibration Sensing, CVPR2022.

[3] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, Hao Li. EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation. CVPR2022.

[4] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, Pratul P. Srinivasan. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. CVPR 2022.

[5] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. CVPR2022.

[6] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, Zhenan Sun. AnyFace: Free-style Text-to-Face Synthesis and Manipulation. CVPR2022.

[7] Lingyan Ruan, Bin Chen, Jizhou Li, Miu-Ling Lam. Learning to Deblur Using Light Field Generated and Real Defocus Images. CVPR 2022.

李俊

南京理工大学计算机科学与工程学院教授，国家高层次青年人才计划入选者，江苏省高层次人才计划入选者。主要研究方向为计算机视觉、机器学习、深度学习与计算物理/化学交叉等。Email:junli@njust.edu.cn

西北工业大学戴玉超教授访谈

2022 年 9 月 5 日，《CCF-CV 专委简报》在线采访了西北工业大学博士生导师戴玉超教授。下面是采访实录。

戴老师，您好！首先，请您分享一下您的个人学习和研究经历。

我于 2001 年进入西北工业大学教育实验学院学习，2005 年免试录取西北工业大学电子信息学院硕士研究生，师从何明一教授，2007 年提前攻读博士学位，2008 年至 2009 年受国家留学基金委资助赴澳大利亚国立大学联合培养，师从国际多视角几何奠基者澳大利亚科学院院士 Richard Hartley 教授和李宏东教授，2012 年获得西北工业大学博士学位。2012 年至 2014 年在澳大利亚国立大学计算机研究院从事博士后研究并获得澳大利亚国家基金委 ARC DECRA 人才项目（澳大利亚“优青”），2014 年至 2017 年在澳大利亚国立大学工程研究院担任 ARC DECRA Fellow。2017 年入选国家青年人才计划并入职西北工业大学电子信息学院任教授、博士生导师，2019 年起担任陕西省信息获取与处理重点实验室主任。

您于 2014—2017 年期间在澳大利亚担任 ARC DECRA Fellow，请问是什么原因让您选择了回国，并选择在西北工业大学工作呢？

我本人自 2012 年博士毕业之后在澳大利亚国立大学从事博士后研究，获得澳大利亚国家基金委 ARC DECRA 人才项目。在获得 CVPR2012 年最佳论文奖以及 ARC DECRA Fellow 入职澳大利亚国立大学工程研究院之后，的确有多所国外名校和国内 985 大学向我伸出橄榄枝希望加盟。

我的学士、硕士和博士学位都是在西北工业大学获得的，从 2005 年本科毕设到 2012 年博士毕业，一直师从何明一教授耕耘在多视角三维重建这一前沿方向，期间获得 IEEE CVPR 2012 最佳论文奖。母校西北工业大学和导师何明一教授的培养在我的心里留下了永恒的烙印。博士毕业后在澳大利亚工作期间，仍然保持与何明一教授及其课题组开展深入的科研合作和人才培养工作，参与何老师牵头申报并获批的国家自然科学基金国际合作重点项目以及国家留学基金委创新型人才培养国际项目等。2016 年依托西北工业大学电子信息学院申请并获批 2017 年度国家级青年人才项目。回国加盟母校西北工业大学电子信息学院，以自己所学进一步培养高素质拔尖人才就是一个非常自然的选择。

您获得了 IEEE CVPR 2012 最佳论文奖（大陆高校 30 年来首次获得该奖项），能跟大家分享一下您获得此奖项的成果及经历么？

我们的获奖论文工作是关于复杂动态场景的单目三维重建，国际上 1992 年提出了随着时间外形可变化的非刚性目标三维重建问题，但由于其欠定性和多解性成为难题。2000 年美国斯坦福大学学者对目标和观测相机给出多个先验条件下的求解方法并获得 IEEE CVPR 最佳论文奖。多先验方法一般假设目标具有特定

运动特性，然而这些假设在现实应用中往往不可能满足。针对以上问题，我们发现了复杂动态场景的本质特性，即刚性与非刚性目标的运动平滑性和三维重建矩阵的低秩性，在此基础上抛弃了二十多年来已有方法使用的多重先验，形成了利用问题本质特性的无先验非刚性目标三维重建理论与方法，不但重建精度显著提升，而且方法简单有效，计算量显著减少，被 IEEE CVPR 2012 技术委员会评价为重大理论突破并授予最佳论文奖（该项获奖是设立 30 余年来中国大陆高校首次获此奖项）。我们的后续工作进一步拓展到非刚性物体稠密三维重建、多非刚性物体三维重建、动态场景稠密三维重建等，形成复杂动态三维重建体系，并获得 IEEE CVPR 2017 非刚性重建挑战赛最佳算法奖等奖项。当前，我们继续在该方向深入研究，结合深度学习等数据驱动模型方法，致力于破解复杂多样现实条件下的复杂动态场景高精度快速稠密重建难题。

您在复杂动态场景的三维重建与感知、深度学习和几何模型融合的稠密匹配、新型仿生视觉传感器和计算成像等方面做出了突出贡献。能跟大家介绍一下您认为最值得骄傲的一项或几项成果么？

我们课题组专注于机器视觉与智能处理领域的科学研究，聚焦于复杂动态场景的三维重建与感知、融合深度学习与多视角几何的三维重建、新型视觉传感器建模与应用、视觉显著性检测等方面。除了上面提到的复杂动态场景三维重建，我这里介绍下我们近来围绕融合深度学习与多视角几何的三维重建和新型视觉传感器建模与应用方面的工作。

在融合深度学习与多视角几何的三维重建方面，针对如何有效结合几何模型与深度学习模型实现无监督学习问题，提出新颖的深度递归神经网络，首创具有时序约束的无监督双目深度感知方法，面向开放环境、陌生环境等。针对如何以多视角几何的恰定性破解深度几何模型的欠定性问题，提出融入多视角几何领域知识的深度结构与运动恢复方法，运动轨迹恢复结果和深度图恢复结果得到显著提升。以上方法在国际自动驾驶数据集长期排名前列，获 ECCV 2020 鲁棒视觉挑战赛双目立体匹配赛道冠军和光流估计赛道亚军、IEEE CVPR 2021 双目立体匹配冠军。

在新型视觉传感器建模与应用方面，针对全局快门构建的三维视觉模型无法应用于卷帘快门相机、限制 CMOS 传感器三维视觉应用的难题，我们统一卷帘快门极线几何模型体系，填补理论空白，同时攻克从卷帘快门图像恢复全局快门图像及高帧率视频的瓶颈难题，致力于推动 CMOS 传感器的广泛三维视觉应用。面向高动态、高速等极端视觉条件应用，研究以事件相机为代表的仿生视觉传感器，建立事件相机成像的二次积分模型，高帧率重建问题简化为单一变量优化问题，极具模型简洁与普适性。课题组也在进一步拓展仿生视觉传感器在航空航天等极端视觉环境下的应用。

除了 IEEE CVPR 2012 最佳论文奖，您还获得了火箭军“智箭火眼”人工智能挑战赛“珠联璧合”科目全国第一名、IEEE CVPR 2020 最佳论文提名奖等，请问您是如何做到在获奖方面持续产出的？

以我们在视觉显著目标检测与伪装目标检测方面的工作为例，我们提出以不确定性建模刻画目标标注中的不确定性，将已有方法的点估计建模提升到分布建模问题，从而有效解决了 RGB-D 设置下显著目标检测难题。此项研究工作获得 CVPR 2020 最佳论文奖提名（Best Paper Nominee）。进一步，我们同时获取显著目标检测与伪装目标检测以及对应的不确定性，在国际上首先建立了显著目标检测与伪装目标检测之间的对立统一框架，实现两项任务的相互促进。

简而言之，我们围绕领域内核心问题，从问题建模、求解思路和现实应用引导等不同方面开展工作，同时紧密结合学生研究工作方向和相关的竞赛活动，实现两者之间的相互促进，实现“以赛促教”和“以赛促学”。

您主持国家自然科学基金面上项目等科研项目，能否跟大家分享一下您在承担这些项目过程中所获得的经验和认识？

西北工业大学同时在航空、航天、航海等领域开展人才培养和科学研究工作，为我们课题组的科研工作开展提供很好的应用背景。我们所从事的机器视觉与人工智能方向主要围绕航空航天航海等对于无人系统自主感知与认知能力的迫切需求开展，承担的科研项目也是围绕以上方向，致力于赋予无人系统感知和认知复杂动态三维环境的能力。课题组开展研究工作努力实现面向世界科技前沿与面向国家重大需求之间的平衡兼顾，努力为国家科技自立自强贡献自己的力量。

您接收研究生的条件是什么？您又是如何对他们进行指导和管理的呢？

在接收研究生方面，我们注重考察学生的综合素质和发展潜力，一般有一定时间的持续沟通和深入交流，从而对于学生有一个全面的考察，同时学生也可以全面了解我们课题组的人才培养模式。

在研究生指导和管理中，我们努力做到因材施教，培养学生科研兴趣和综合能力。在研究生人才培养中，不仅注重培养学生如何做研究、写论文，而且注重培养学生思想品格以及对科研的兴趣、习惯和品味。对于实验室博士生和硕士生，摸清学生个人兴趣及其能力特点，注重根据学生的差异实现个性化的培养，结合学术前沿和工程应用需求，安排恰当的研究课题，充分调动其学习积极性。对于具有较强学术潜力的研究生，鼓励大胆、自由的学术探索，并通过及时交流沟通，保证学术方向的正确性。在研究生指导过程中，秉承科学理念，注重培养理论联系实际的科研风格，将研究兴趣和国家需求相结合，敢于挑战开放性问题和迎战技术难题，实验室形成了良好的学术和科研氛围。指导的研究生在领域内顶级国际会议 IEEE CVPR、IEEE ICCV 等发表高水平论文，一名博士生入选 CVPR 2022 博士论坛（该年度入选者中的唯一中国高校在读博士生）。

如果吐露研究工作者的心声，您最想说的是什么？

紧密围绕领域核心问题，不忘科研初心，持续深耕，终会拨云见日。大家一起共同努力，持续扩大中国在计算机视觉领域的国际影响。

责任编委 赵振兵 余烨

戴玉超

戴玉超，男，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、人工智能等领域，聚焦复杂动态场景的三维重建与感知、深度学习和几何模型融合的稠密匹配、新型仿生视觉传感器和计算成像等问题。主持国家自然科学基金面上项目、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 IEEE 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 6300 次，H 因子 38。先后获得 IEEE CVPR 2012 最佳论文奖（大陆高校 30 年来首次获得该奖项）、火箭军“智箭火眼”人工智能挑战赛全国第一名、IEEE CVPR 2020 最佳论文奖提名、ECCV 2020 鲁棒计算机视觉挑战赛双目深度估计赛道冠军和光流估计赛道亚军、2014 DICTA DSTO 图像处理最佳基础贡献论文奖、CVPR 2017 非刚性结构与运动恢复挑战赛最佳算法奖、DICTA 2017 最佳学生论文奖、APSIPA 2017 年度峰会最佳深度学习/机器学习论文奖、陕西省优秀博士论文和陕西省科技进步二等奖等奖项。担任 APSIPA 杰出讲者和 IEEE CVPR、IEEE ICCV、ACM MM 等国际顶级会议领域主席，ACCV 2022 宣传主席，中国图象图形学报青年编委。

委员好消息

✪ 2022 年 6 月 23 日，2022 IAPR Fellow 名单公布，CCF-CV 专委会 5 位执行委员当选：西北工业大学**韩军伟**因在视觉显著性计算和遥感影像分析方面的贡献、复旦大学**姜育刚**因在大规模和可信视频理解以及开源数据集方面的贡献、中科院自动化所**雷震**因在人脸分析和模式识别方面的贡献、中山大学**林倞**因在面向视觉模式匹配和理解的大规模学习算法与模型方面的贡献、南京理工大学**唐金辉**因在多媒体内容分析与识别方面的贡献当选 2022 年度 IAPR Fellow。

✪ 2022 年 7 月 11 日，2021 年度浙江省科学技术奖获奖名单公布，CCF-CV 专委会执行委员、浙江大学**章国锋**和**周晓巍**参与完成的“单目视觉鲁棒跟踪定位的理论和方法”获自然科学一等奖。

✪ 2022 年 8 月 9 日，2022 年度 IAPR Award 获奖者名单公布，CCF-CV 专委顾问委员会主任、中国科学院自动化研究所**谭铁牛**获 King-Sun Fu 奖，CCF-CV 专委顾问委员会委员、北京航空航天大学**王蕴红**获 Maria Petrou 奖。

✪ 2022 年 8 月 12 日，2021 年度北京市科学技术奖评审委员会评审结果公示，CCF-CV 专委会副秘书长、北京工业大学**毋立芳**、CCF-CV 专委会执行委员、北京工业大学**简萌**等完成的“多光源可调节的面曝光 3D 打印关键技术及应用”获技术发明二等奖，CCF-CV 专委会执行委员、中科院自动化研所**赫然**和 CCF-CV 顾问委员会主任、中科院自动化研究所**谭铁牛**院士等完成的“视觉内容智能解析与合成的关键技术和应用平台”获科技进步二等奖，CCF-CV 专委会执行委员、北京航空航天大学**徐迈**参与完成的“多媒体计算通信技术与智能安防系统研发及应用”获科技进步一等奖。

✪ 2022 年 8 月 22 日，中国科学院公布了 2022 年度“中国科学院优秀导师”名单，共 173 人获此称号。CCF-CV 专委会 3 位执行委员上榜，他们是：中国科学院大学**黄庆明**（中科院信工所客座教授）、中国科学院自动化研究所**赫然**、中国科学院沈阳自动化研究所**丛杨**。

✪ 2022 年 6 月 27 日，黑龙江省教育厅下发了《省教育厅关于公布 2022 年黑龙江省高等教育和职业教育教学成果奖获奖名单的通知》，公布了 2022 年黑龙江省高等教育和职业教育教学成果奖评审结果，确定高等教育教学成果奖 260 项，CCF-CV 专委会执行委员、哈尔滨工程大学**刘海波**参与完成的“一流引领，认证保障，计算机类创新人才培养模式建设与实践”获二等奖。

✪ 2022 年 9 月 17 日，2022 年度中国人工智能学会会士名单公示，共 13 人入围，CCF-CV 专委会顾问委员、北京航空航天大学**王蕴红**上榜。

✪ 2022 年 9 月 20 日，上海市优秀教学成果项目名单公示，高等教育优秀教学成果项目共 532 项，CCF-CV 专委会 4 位执行委员的 5 项教学成果入围：上海科技大学**虞晶怡**参与完成的“‘全方位、全流程’研究创新型本科人才培养的探索与实践”拟授本科教育教学成果特等奖，上海科技大学**高盛华**、**虞晶怡**等完成的“基础与前沿融合、理论与实践并进——服务国家发展战略的计算机视觉教学体系实践”拟授本科教育教学成果二等奖，同济大学**张林**参与完成的“筑基建魂，二十年砥砺求索，新时代一流软件工程人才”拟授本科教育教学成果二等奖，上海大学**曾丹**参与完成的“全科全所，同频共振：科教融合打造长三角高层次创新人才培养高地”拟授研究生教育教学成果二等奖。

责任编委 刘海波

人脸属性合成领域开源代码

华北电力大学　王艺博 张珂

人脸属性合成是指对原始人脸的某些属性进行编辑，合成目标属性的人脸图像。目前较为主流的方法是利用 GAN 及其他深度生成模型对肤色、年龄、眼镜、胡须、发色、发型等面部特定特征进行修改。其难点在于合成图像仅需改变与目标属性相关的人脸区域，其他无关属性保持不变。人脸属性合成由于其重要的应用价值，已成为计算机视觉领域的重要研究方向之一。本文重点介绍一些人脸属性合成领域的研究成果。

1、StarGAN v2

工作： StarGAN 是最早的图像翻译模型之一，它使用单个生成器学习所有可用域之间的映射。生成器将域标签作为附加输入，并学习将图像变换为相应的域。然而，由于每个域由预定标签指示，StarGAN 只能学习每个域的确定性映射，不能捕获数据分布的多模态特性。

该文提出了一种可扩展的方法 StarGAN v2，该方法可以跨多个域生成不同的图像。假设 X 和 Y 分别为原始图像集和目标域，给定图像 x∈X 和任意域 y∈Y，StarGAN v2 的目标是训练生成器生成与图像 x 对应的每个域 y 的不同风格的图像，即在每个域的学习样式空间中生成目标域的样式向量，并训练生成器学习样式向量的表示形式。

StarGAN v2 将 StarGAN 的域标签替换为特定域的样式码。StarGAN v2 的网络结构图如图 1 所示，由生成器、映射网络、风格编码器与鉴别器组成。映射网络学习将随机高斯噪声变换为样式码，风格编码器学习从给定参考图像提取样式码。由于网络涉及多个域，两个模块均设计了多个输出分支，每个分支为特定域提供样式码。最后，生成器利用这些样式代码，学习如何在多个域上成功地生成不同属性的图像。

StarGAN v2 的损失函数包括对抗性损失、风格重建损失、多样性敏感损失以及循环一致性损失。其中，风格重建损失强制生成器在生成图像时使用样式码；多样性敏感损失使生成器生成风格多样化的图像；循环一致性损失强制生成的图像保留其对应输入图像的域不变特征（如姿态）。

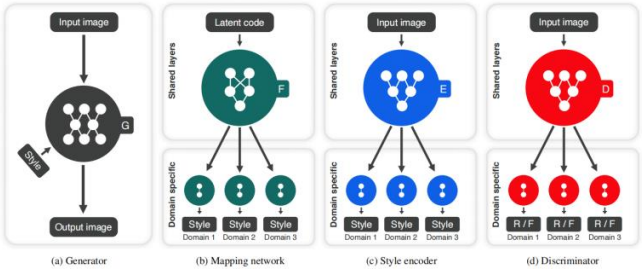


图 1 StarGAN v2 结构图

更多有关 StarGAN v2 的详细内容可参考发布该方法的论文“StarGAN v2: Diverse Image Synthesis for Multiple Domains”。

论文地址： https://arxiv.org/pdf/1912.01865v2.pdf
代码地址： https://github.com/clovaai/stargan-v2

2、Multi-attention U-Net-based Generative Adversarial Network（MU-GAN）

工作： 该文提出了一种基于 U-Net 的多级视觉注意生成对抗网络模型 MU-GAN。MU-GAN 能够生成具有更

高视觉保真度、属性操作准确性和几何合理性的面部图像。

为了合成目标图像时保留更多面部细节，生成器采用了对称 U-Net 架构，有效避免了由于最后解码器层的信道数急剧减少而导致的信息丢失。MU-GAN 将自注意力模块作为卷积层的补充引入编码器-解码器架构中，其结构图如图 2 所示。

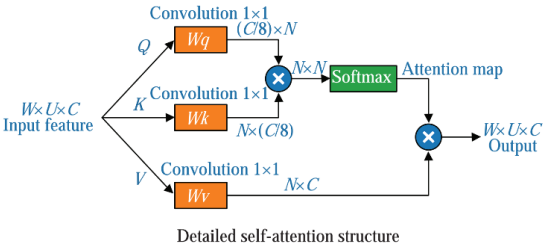


图 2 自注意力模块结构图

MU-GAN 的损失函数包括对抗性损失、属性分类损失以及重建损失。为了精确地将图像转换为具有目标面部属性的图像，MU-GAN 使用属性分类器对面部属性进行分类，其通过属性分类损失对生成具有正确面部属性的图像施加属性约束。使用对抗性损失和属性分类损失不能保证网络只改变目标属性相关区域。因此，生成器需要在原始属性标签条件下学习从潜空间中重建图像。重建损失引入 L1 范数来衡量生成图像与输入图像之间的相似性。

该文在 CelebA 数据集上进行了大量实验，这些实验证明了 MU-GAN 在人脸属性合成任务中的有效性和潜力。 MU-GAN 的网络结构图如图 3 所示

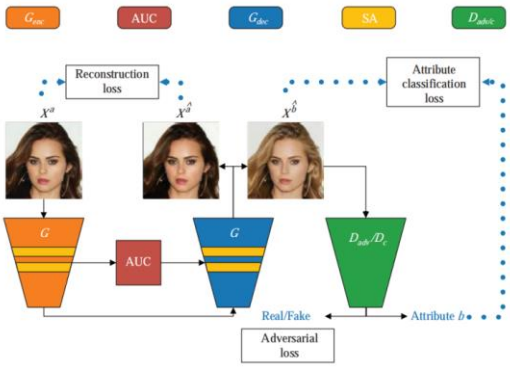


图 3 MU-GAN 结构图

更多有关 MU-GAN 的详细内容可参考发布该方法的论文“MU-GAN: Facial Attribute Editing Based on Multi-Attention Mechanism”。

论文地址： https://ieeexplore.ieee.org/document/9 205685
代码地址： https://github.com/SuSir1996/MU-GAN

3、Hierarchical Style Disentanglement（HiSD）

工作： 该文提出了一种层次风格解耦模型 HiSD ，通过将原始标签设计为一个层次结构来解决多标签和多样式图像转换中的问题，该层次结构从上到下分为独立标签、排他属性和分离样式三部分。HiSD 将标签在图像中的表现形式视作标签相关样式，并标识为标签和属性两方面，这种方法为合成不同属性的图像提供了更加可控的方向，其原理如图 4 所示。

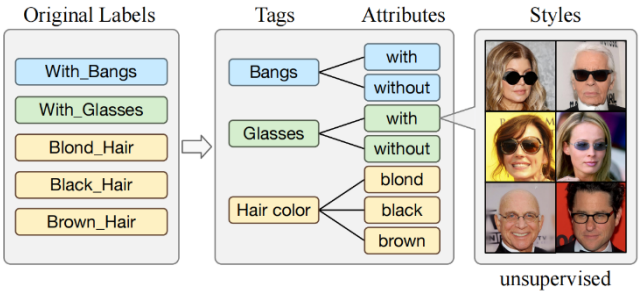


图 4 HiSD 原理图

HiSD 通过引入不同的模块来生成、提取和有效地操作解耦的标签相关样式。在循环平移路径中，模型始终优化生成、提取的样式，以真实准确地操作图像。对于无监督的风格解耦，HiSD 引入了两种架构进行改进，局部翻译器使用注意力掩码来避免全局操作，无关标签条件鉴别器在注释中使用冗余标签以防止隐式属性被转换。HiSD 的网络结构图如图 5 所示。为了独立地优化不同标签及属性的模块，HiSD 在每次训练过程中对标签、源属性和目标属性随机地进行采样，其训练阶段包括三个路径，即非翻译路径、自翻译路径与循环翻译路径。

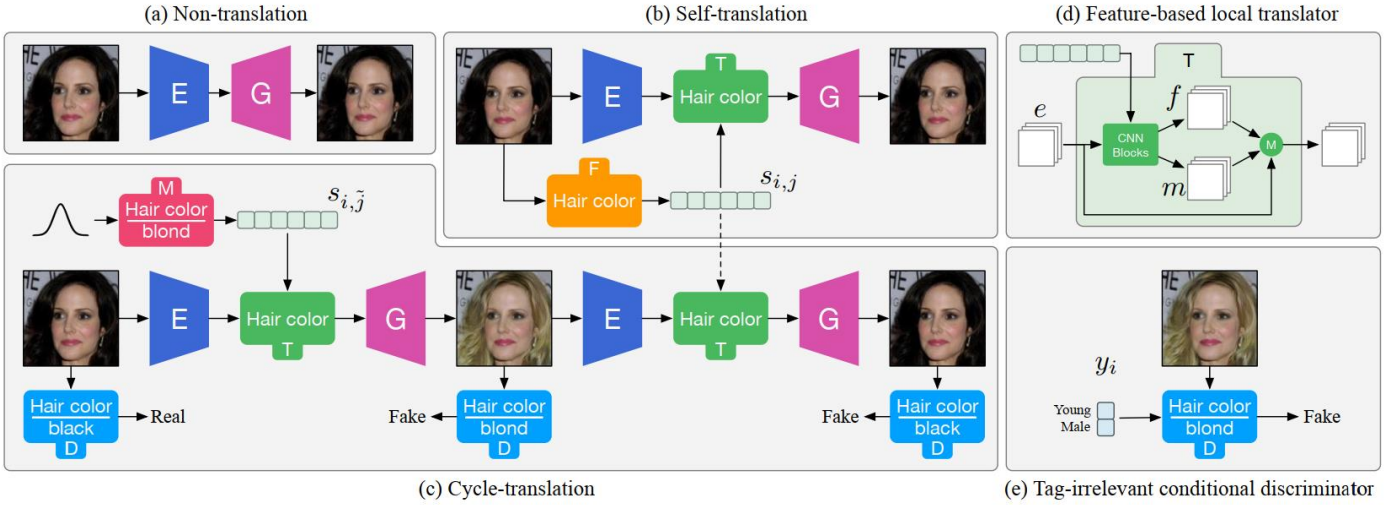


图 5 HiSD 结构图

HiSD 的损失函数包括对抗性损失、重建损失以及风格损失。对抗性损失鼓励网络有效地生成和提取标签相关样式。由于非翻译、自翻译与循环翻译路径的输出都是输入图像的重建图像，因此，HiSD 应用重建损失使重建图像与输入图像尽可能相同。风格损失保证生成图像和输入图像之间风格的一致性，它鼓励映射器生成准确的标签相关样式码，翻译器充分利用该样式码。

更多有关 HiSD 的详细内容可参考发布该方法的论文“Image-to-image Translation via Hierarchical Style Disentanglement”。

论文地址： https://arxiv.org/pdf/2103.01456.pdf

代码地址： https://github.com/imlixinyang/HiSD

责任编委 **李策 樊鑫**

王艺博

硕士研究生，华北电力大学电子与通信工程系，研究方向为人脸属性合成。

张珂

教授，博士生导师，华北电力大学电子与通信工程系从事教学与科研工作。研究方向为计算机视觉、人脸属性分析、电力计算机视觉。

个人主页：

https://dece.ncepu.edu.cn/szdw/xxcljys/xxcljys_js/8cacc0a0cd2d41ada805039e75d4b35d.htm

行为动作识别数据集

西安电子科技大学 李洪升 张亮

随着在线媒体、监控和移动摄像头的增长，视频数据库的数量和规模正以惊人的速度增长。动作识别任务已经成为当前研究的热点问题，相比图像来说，视频内容和背景更加复杂多变，不同的动作类别之间具有相似性，而相同的类别在不同环境下又有着不同的特点。然而，尽管视频数据呈爆炸式增长，但自动识别和理解人类活动的能力仍然相当有限。其中一个关键的因素在于可用的动作识别数据集。

与图像识别相比，动作识别需要更大量数据。目前，已经出现不同规模的动作识别数据集，虽然相比实际情况仍不够全面，但也对动作识别的研究有了极大的帮助。

ADNI 的数据目前分为四个阶段,ADNI-GO、ADNI-1、ADNI-2 和 ADNI-3,其中 ADNI-GO 与 ADNI-1 为基线数据,ADNI-2 与 ADNI-3 主要为后续跟踪数据和新加入的模态数据(如新型示踪剂下的 PET)等。

2、动作识别数据集简介

动作识别数据集与图像数据集类似，经历了一个数据集量由少到多，标注由简单到详尽的过程。在本章中，将按照时间的顺序分别介绍 HMDB51、UCF101、ActivityNet 、 Charades 、 Kinetics 、 Something-Something 共六类常用的动作识别数据集。

HMDB-51

HMDB-51 (a Large Human Motion DataBase) 动作识别数据集，主要收集自电影数据，其余一小部分来自其他公共数据集，如 Prelinger 档案、YouTube 和 Google 视频。该数据集包含 6,849 个片段，分为 51 个动作类别，每个类别至少包含 101 个片段。动作类别分为五种类型：

一般面部动作微笑：大笑，咀嚼，说话。

具有对象操纵的面部动作：吸烟、吃、喝。

一般身体动作：侧手翻、拍手、爬、爬楼梯、潜水、倒地、反手翻转、倒立、跳跃、引体向上、俯卧撑、跑、坐下、坐起来、翻筋斗、站起来、转身、走，海浪。

与物体互动的身体动作：刷头发、接球、拔剑、运球、打高尔夫球、打东西、踢球、捡、倒、推东西、骑自行车、骑马、射球、射弓、射枪、挥动棒球棒、剑术，投掷。

人类互动的身体动作：击剑、拥抱、踢某人、亲吻、拳击、握手、剑斗。

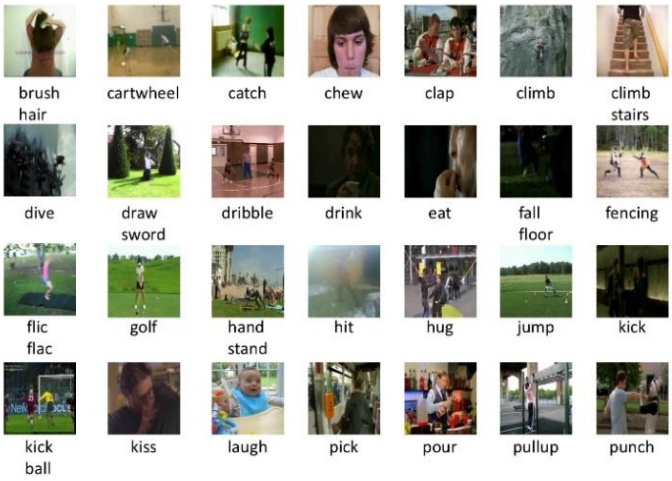


图 1 HMDB-51 数据集样本示例

数据集地址：

https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/

UCF101

UCF101(University of Central Florida)是一个由真实动作视频组成的动作识别数据集，视频数据收集自YouTube，总共包含 101 个动作类别。该数据集为其前身 UCF50 数据集的扩展，具有 50 个动作类别。

UCF101 拥有来自 101 个动作类别的 13,,320 个视频，平均每个类别包含约 129 个视频片段，在动作方面具有不同程度的多样性，且在相机运动、物体外观和姿势、物体尺度、视点、杂乱背景、照明条件等方面存在很大变化。与其他通过演员表演的动作不同，UCF101 动作识别数据集旨在收集真实的显示数据以保证充分的动作多样性。

UCF101 数据集中的 101 个动作类别中的视频分为 25 组，每组可以包含 4-7 个动作视频。来自同一组的视频可能具有一些共同的特征，例如相似的背景、相似的视角等。动作类别总共可分为五种类型：1）人与物体互动 2）仅身体运动 3）人与人互动 4）演奏乐器 5）运动。

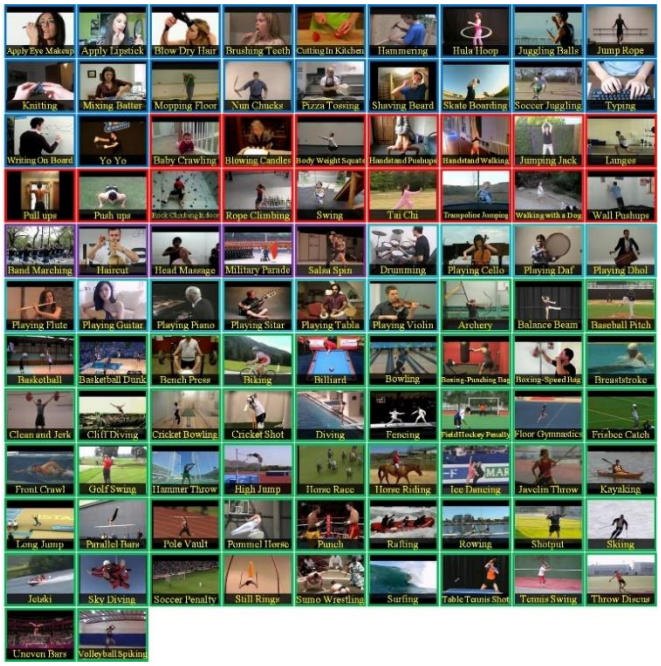


图 2 UCF101 数据集样本示例

数据集地址：

https://www.crcv.ucf.edu/data/UCF101.php

ActivityNet

ActivityNet从ATUS(American Time Use Survey,美国时间使用调查)提供的两千多个活动中手动选择了203 个活动类别的子集。这些活动属于 7 个不同的顶级类别：个人护理、饮食、家庭、关怀和帮助、工作、社交和休闲以及运动和锻炼。

ActivityNet 提供了 203 个活动类别的样本，每个类平均有 137 个未修剪的视频，总共有 20K 多个 Youtube 视频，其中训练包括 10K 多个视频，验证与测试各包含 5K 个视频。每个视频平均有 1.41 个行为标注，共计 849 个小时视频时长。

所有 ActivityNet 视频均来自在线视频共享网站。大部分视频的时长在 5 到 10 分钟之间。大约 50%的视频是高清分辨率(1280x720)，且大多数视频的帧速率为 30FPS。

图 3 ActivityNet 数据集类别层次

数据集地址：

http://activity-net.org/

Charades 数据集

Charades 是一个大规模数据集，重点关注使用 Holly-wood-in-Homes 方法收集的常见家庭活动。Charades 数据集从 Amazon-Mechanical-Turk 招募了数百人来表演预先定义的动作。工作人员还提供动作分类、本地化和视频描述注释。

Charades 数据集的第包含 9848 个日常活动视频，平均时长为 30.1 秒。其中训练集包含 7,985 个片段，测试集包含 1,863 个片段。该数据集收集了 15 种类型的室内场景，涉及与 46 个对象类的交互，并具有 30 个动词的词汇表，这些动作总共有 157 个动作类。它有 66,500 个时序定位动作，平均时长 12.8 秒，由三大洲的共 267 人进行录制。

图 4 Charades 数据集部分类别样本示例

数据集地址：

https://prior.allenai.org/projects/charades

Kinetics 数据集

Kinetics 数据集目前总共发布了三个版本，分别为 Kinetics-400，Kinetics-600 和 Kinetics-700。这三个数据集在样本数量和类别数量上都进行了增加。

Kinetics-400

该数据集专注于人类行为（而不是活动或事件）。动作类列表包括：人动作（单数），例如画画、喝酒、大笑、抽拳；人对人的行动，例如拥抱、亲吻、握手；以及人-对象动作，例如打开礼物，修剪草坪，洗碗。有些动作需要时间推理来区分，例如不同类型的游泳。其他动作则需要更加强调对象来区分，例如演奏不同类型的管乐器。

没有很深的层次结构，而是有几个（非排他性的）父子分组，例如音乐（打鼓、长号、小提琴……）；个人卫生（刷牙、剪指甲、洗手……）；跳舞（芭蕾、玛卡丽娜、踢踏舞……）；烹饪（切割、油炸、去皮……）。

该数据集有 400 个人类动作类，每个动作有 400-1,150 个片段，每个片段来自一个单独的视频。每个片段持续约 10 秒。Kinetics-400 总共有 306,245 个视频，分为三个部分，一个用于训练，每个类别有 250-1000 个视频，一个用于验证，每个类别有 50 个视频，一个用于测试，每个类别有 100 个视频。这些片段来自 YouTube 视频，具有不同的分辨率和帧速率。

每个类都包含说明该动作的片段。但是，一个特定的片段可以包含多个动作。数据集中有趣的例子包括：“开车”时“发短信”；“弹尤克里里”的同时“呼啦圈”；“跳舞”（某种类型）时“刷牙”。在每种情况下，这两个动作都是 Kinetics 类，并且片段可能只会出现在其中一个类，即片段没有完整（详尽的）注释。

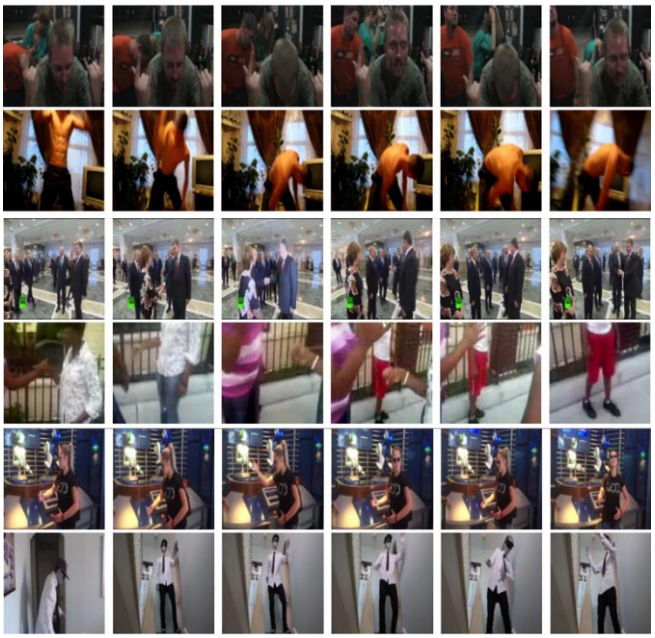


图 5 Kinetics-400 数据集部分类别样本示例

Kinetics-600

新版本的数据集称为 Kinetics-600，遵循与 Kinetics-400 相同的原则：

(i) 片段来自 YouTube 视频，每个片段持续 10 秒，具有不同的分辨率和帧速率；

(ii) 对于动作类，所有片段都来自不同的 YouTube 视频。Kinetics-600 代表课程数量增加了 50%，从 400 增加到 600，视频片段数量增加了 60%，从大约 300k 增加到大约 500k。两个数据集版本的统计数据详见表 1。在新的 Kinetics-600 数据集中，有一个标准测试集，其标签已公开发布，还有一个保留测试集（标签未发布）。

Kinetics-700

新数据集遵循与 Kinetics-400 和 Kinetics-600 相同的原则：

(i) 片段来自 YouTube 视频，每个片段持续 10 秒，具有可变的分辨率和帧速率；

(ii) 对于动作类，所有片段都来自不同的 YouTube 视频。Kinetics-700 几乎是 Kinetics-600 的超集：类的数量从 600 个增加到 700 个，保留了 Kinetics-600 中的三个类以外的所有类。与 Kinetics-600 的情况一样，Kinetics-700 每个人类动作类别有 600 个或更多片段，视频片段的数量增加了 30%，从大约 500k 增加到大约 650k。

图 6 Kinetics-700 数据集部分类别样本示例

数据集地址：

https://www.deepmind.com/open-source/kinetics

Something-Something

Something-Something 数据集是一个更加细粒化的数据集，该数据集主要关注的是动作本身而非动作相关的对象，即当同一种动作作用与不同的对象时，在 something-something 数据集中被归类为一类动作。其中 v2 是 v1 版本的扩充，因此目前更常用 v2 版本，v1 版本基本已经弃用。

Something-Something v1

20BN-Something-Something 数据集是大量标记视频片段的集合，这些片段显示人类使用日常物体执行预定义的基本动作。该数据集是由大量工作者创建的。它允许机器学习模型对物理世界中发生的基本动作进行细致地理解。它包含 108,499 个视频，其中 86,017 个在训练集中，11,522 个在验证集中，10,960 个在测试集中。有 174 个标签。时长从 2 秒到 6 秒不等。标签是基于模板的文本描述，例如“Dropping [something] into [something]”，其中包含用作对象占位符的槽（“[something]”）。工作者提供他们表演模板的视频。他们选择要对其执行操作的对象，并在上传视频时输入描述对象的名词短语。

数据集以 8:1:1 的比例分为训练集、验证集和测试集。创建拆分是为了确保同一工作人员提供的所有视频仅在一次拆分（训练、验证或测试）中出现。包括大小写、词干、限定词的使用等方面的差异，在当前版本的数据集中提交了 23,137 个不同的对象名称。在 V1 版本中，该数据集由 1133 名工作者生成，平均每个类别由 127.32 名工作者完成。

20BN-Something-Something V2

V2 数据集实在 V1 数据集上进行的增广。该数据集是同样由大量工作者创建的。它包含 220,847 个视频，其中 168,913 个在训练集中，24,777 个在验证集中，27,157 个在测试集中。有 174 个标签。

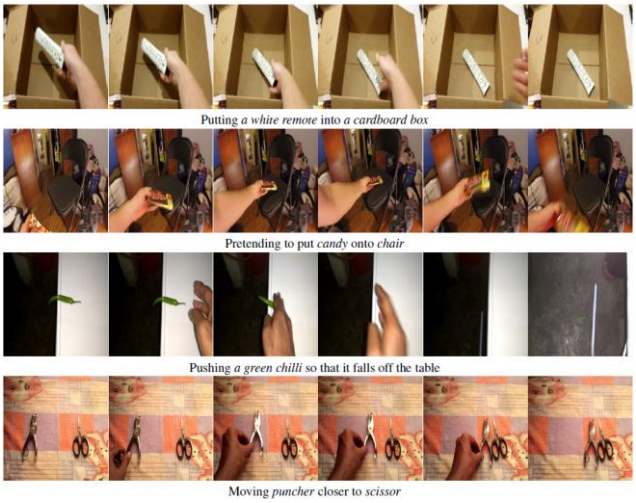


图 7 Something-Something V1 数据集部分类别样本示例

数据集地址：

https://developer.qualcomm.com/software/ai-

1、动作识别数据集简介

从表 1 中可以看出，在当前常见的 HMDB51、UCF101 、 ActivityNet 、 Charades 、 Kinetics 、 Something-Something 这 6 类数据集中，Kinetics-700 类别数最多，样本量最丰富。基本接近图像识别领域的 ImageNet 水平（ImageNet 包含 1,000 类，共 14,197,122 个样本），可以为其他视频任务提供合适的迁移模型。但同时，庞大的数据量也为模型的训练增加了巨大的时间成本。与 Kinetics-700 相比，Something-Something V2 数据集，平均每个类别拥有 1.27K 个样本，且拥有合适的数据量（221K），是用来做动作识别任务的相对合适的数据集.

表 1 现有的常见数据集属性对比

数据集	动作数	样本数	发布年份
HMDB-51	51	7K	2011
UCF101	101	13K	2012
ActivityNet	200	20K	2015
Charades	157	7k	2016
Kinetics-400	400	300K	2017
Kinetics-600	600	500K	2018
Kinetics-700	700	650K	2019
Sth-V1	174	108K	2017
Sth-V2	174	221K	2017

责任编委 沈沛意 贾同

李洪升

博士研究生，西安电子科技大学计算机科学与技术学院，研究方向为序列数据深度网络建模方法、动作行为识别。

张 亮

教授，博士生导师，西安电子科技大学计算机科学与技术学院从事教学与科研工作。研究方向为深度网络结构设计、自然人机交互、场景感知理解等。
个人主页：https://faculty.xidian.edu.cn/ZL_16/zh_CN/index.htm

好文推荐

南开大学“Re-Thinking Co-Salient Object Detection”的最新成果发表在 IEEE TPAMI 2022。

论文：Deng-Ping Fan , Tengpeng Li, Zheng Lin , Ge-Peng Ji , Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-Thinking Co-Salient Object Detection, IEEE TPAMI, 44(8): 4339-4354 (2022)

显著性检测在 RGB-D 图像、视频，彩色图像等领域有重要研究意义。显著性目标检测是指模拟人类视觉系统，以检测单幅图像中最吸引注意力的物体。共同显著性目标检测是显著性检测方向的一类扩展，其目标是 s 对象的两个重要特征是局部显著性和全局相似性。共同显著性目标检测在农作物采集与感知、协同分割、弱监督学习、图像检索、视频前景检测等领域均得到广泛关注与应用。

文章提出了一种简单有效 CoEGNet，通过以无监督的方式引入协同注意力信息，扩展了最先进的显著性检测模型 EGNet。CoEGNet 结构流程图如图 1 所示。CoEGNet 包含两个独立分支，按照从上到下的顺序，两个分支分别记为顶部分支和底部分支。顶部分支将提取的高层图像特征输入到共同注意力投影模块中，为每个输入图像生成共同注意力图。在底部分支，将每一张图像送入边缘引导的显著性检测网络(EGNet)中，以生成显著性先验图。最后，将两层分支的输出使用元素相乘进行集成，得到最终优化的输出。CoEG-Net 充分利用了之前大规模的显著性目标检测数据集，显著提高了模型的可扩展性和稳定性。

综合基准测试结果表明， CoEGNet 优于 18 个流行的共同显著性目标检测方法。此外，CoEGNe 可产生极具竞争力的视觉可视化结果，成为共同显著性目标检测任务的有效解决方案。

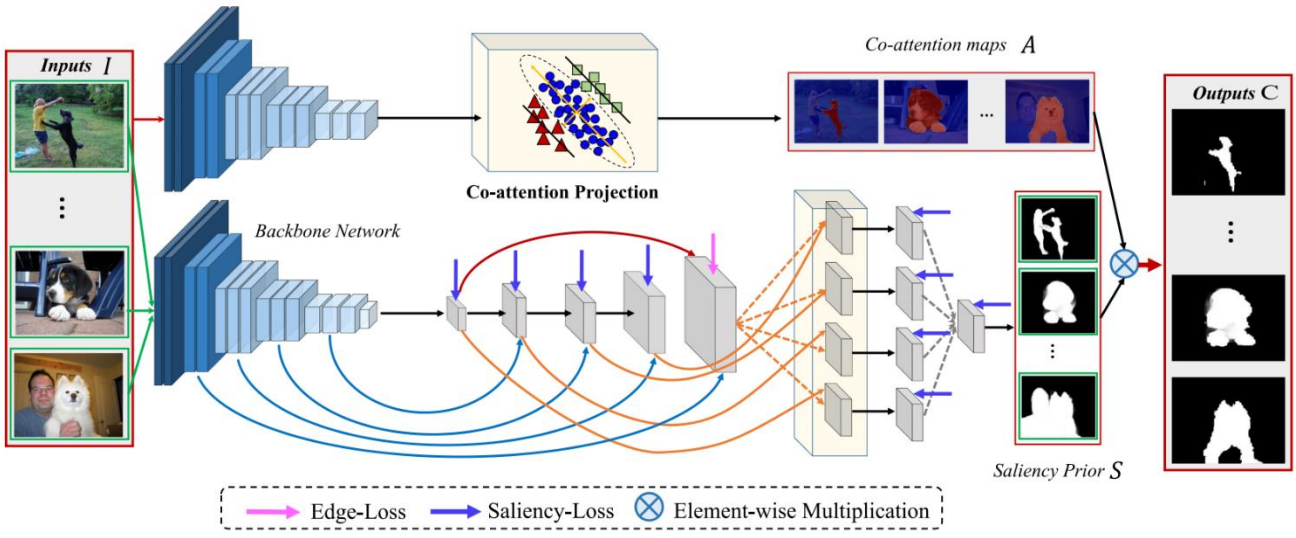


图 1 CoEGNet 模型结构流程图

责任编委 贾同 李策

好文推荐

上海交通大学“Transferable Interactiveness Knowledge for Human-Object Interaction Detection”的最新成果发表在 IEEE TPAMI 2022。

论文：Yong-Lu Li , Xinpeng Liu , Xiaoqian Wu , Xijie Huang , Liang Xu , and Cewu Lu. Transferable Interactiveness Knowledge for Human-Object Interaction Detection, IEEE TPAMI, 44(7): 3870-3882 (2022)

人物交互（Human Object Interaction，HOI）即人-物体交互检测，主要目的是定位人体、物体、并识别他们之间的交互关系。作为视觉关系的子任务，人物交互对行为理解至关重要，可以应用于活动理解，模仿学习等领域。近年来，深度神经网络在人物交互方向取得了令人瞩目的进展。

文章探讨了交互性知识，该知识用于表明人与物之间是否存在交互。研究发现，交互知识可以在人物交互数据集之间学习，并能减小不同人物交互类别设置之间的差距。文章的核心思想是利用交互网络从多个人物交互数据集中学习一般的交互知识。推理过程中，在人物交互分类之前进行非交互抑制。由于交互性的泛化性，交互性网络是一种可迁移的知识学习器，可以与任何人物交互检测模型配合使用，以达到理想的效果。具体而言，文章作者同时利用人体实例和身体部位特征来学习分层范式中的交互性，即实例级和身体部位级的交互性。然后，提出一致性任务来指导学习，以提取更深层的交互视觉线索。文章结构流程如图 1 所示。

文章在 HICO-DET、V-COCO 和新构建的 HAKE-HOI 数据集上广泛评估了所提方法。通过学习交互性，所提方法优于现有的 HOI 检测方法，验证了其有效性和灵活性。

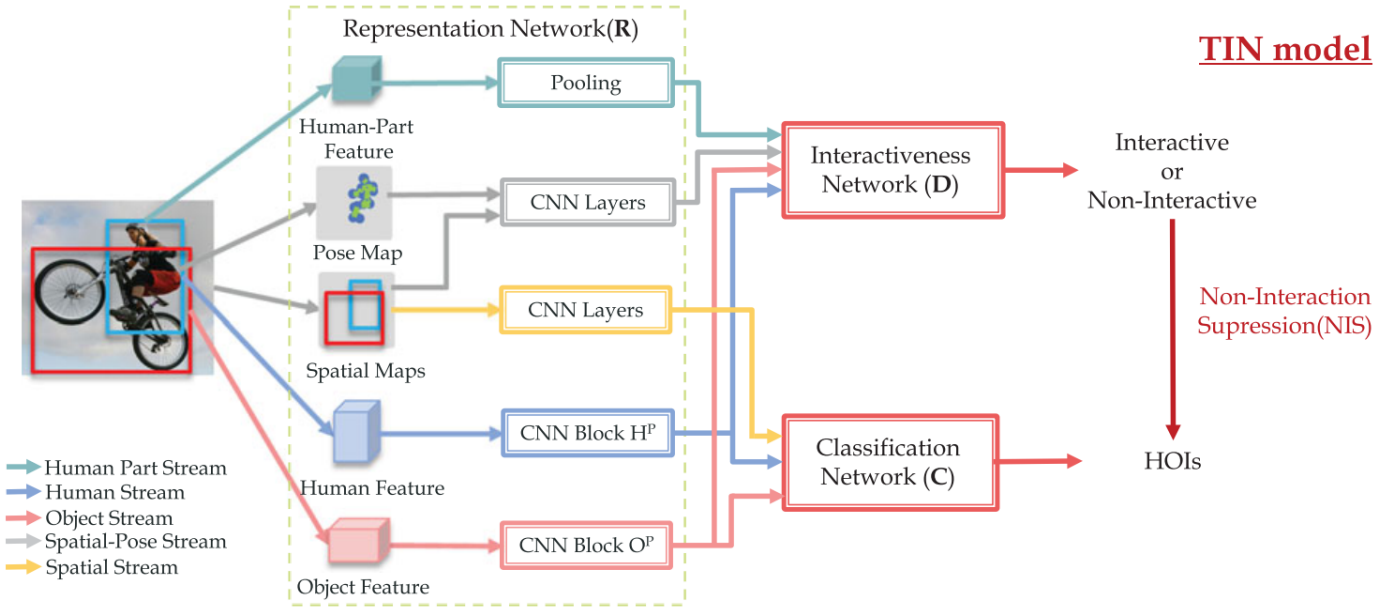


图 1 TIN 模型结构流程图

责任编委 樊鑫 李策

好文推荐

深圳大学团队“生形成物：基于稀疏表示和Transformer的形状生成方法”最新成果发表在CVPR-2022。

论文：Yan X, Lin L, Mitra N J, et al. Shapeformer: Transformer-based shape completion via sparse representation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 6239-6249.

三维物体的数字孪生技术是虚拟现实、计算机辅助设计、机器人等领域的重要基础，数字化的三维物体主要通过两种方式获得：人工建模和扫描重建。为了得到完整、高质量的形状，两者都需花费专业人员大量的时间精力，严重制约了上述领域的快速发展。

如果非专业用户仅利用简单、不完整、可能有较大残缺的低质量输入数据就能得到较好质量的三维形状，那么数字孪生的制作过程就会变得方便快捷，为很多后续应用创造可能性。然而，现有方法在面向低质量输入时均不能取得令人满意的效果。

为了解决上述问题，可以利用生成模型（如：编码器，GANs等）对物体形状的分布进行建模。尽管普通的生成模型对全局信息编码完备，但由于失去了局部空间信息，获取的物体形状往往会丢失很多空间几何结构。如果通过多个局部空间编码组成空间立体块，就可以分化物体整体结构分布，更好的重建出物体三维几何结构，但如何生成这种带有空间信息的复杂表示是一大难题。

为此，深圳大学团队提出了一种基于稀疏表示和Transformer的自回归生成模型：ShapeFormer。该模型可以将不完整形状的序列“翻译”成完整形状的序列，算法流程图如图2所示。通过给定残缺点云数据，向量量化深度隐函数（VQDIF）编码器将其转换为序列表示后，ShapeFormer模型再根据得到的量化序列预测出完整序列的概率分布。最后，对该分布进行采样后解码即可得到多个完整形状，如图1所示。

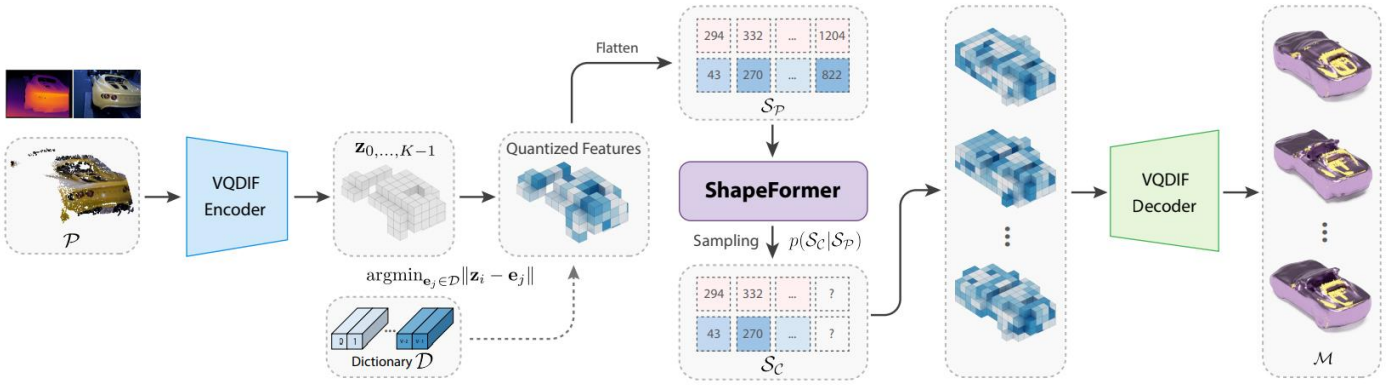


图1 基于稀疏表示和Transformer的形状生成方法的算法流程图

责任编委 贾同 沈沛意

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。

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 Information Sciences, IEEE Transactions on Multimedia, IEEE Signal Processing Magazine, IEEE Journal of Selected Topics in Signal Processing, 和 IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing。

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)，由中国人工智能学会（CAAI）、中国计算机学会（CCF）、中国自动化学会（CAA）和中国图象图形学学会（CSIG）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。

第五届 PRCV 将于 2022 年 11 月 4 日至 7 日在深圳举行，由南方科技大学和深圳职业技术学院共同承办，香港浸会大学、香港中文大学（深圳）、哈尔滨工业大学 (深圳) 和中国科学院深圳先进技术研究所联合承办。本届会议旨为促进 PRCV 和湾区学者交流融合、聚焦前沿理论，提高学术交流氛围和质量、技术赋能产业，吸引科技类企业和投资类企业参与、联合粤港澳，提供一个全国科研团队和粤港澳企业近距离交流科研平台和机会。会议论文集将由 Springer 出版社出版，并被 EI 和 ISTP 检索。

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
WWW 2023	2023.4.30-5.4	Texas, USA	2022.10.14	https://www2023.thewebconf.org/
ICASSP 2023	2023.06.04-7	Rhodes Island, GRC	2022.10.20	https://2023.ieeeicassp.org/
AAMAS 2023	2023.5.29-6.2	London, UK	2022.10.29	https://aamas2023.soton.ac.uk/
CVPR 2023	2023.06.17-23	Vancouver, Canada	2022.11.12	http://cvpr2023.thecvf.com/

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
Information Sciences	Recent Progress in Autonomous Machine Learning	https://www.journals.elsevier.com/information-sciences	2022.10.12
TMM	Point Cloud Processing and Understanding	https://signalprocessingsociety.org/sites/default/files/uploads/special_issues_deadlines/TMM_SI_point_cloud.pdf	2022.10.15
IEEE SPM	Intelligent Signal Processing for Affective Computing	https://signalprocessingsociety.org/sites/default/files/uploads/special_issues_deadlines/SPM_SI_intelligent_signal.pdf	2022.11.01
JSTAR	Cooperative Perception for Computer Vision in Remote Sensing	https://mc.manuscriptcentral.com/jstars	2022.11.30

心底无私视界宽，图像工程视界广

–章毓晋教授专访

本栏目是期望通过计算机视觉及相关领域的前辈回顾个人求学工作经历、回忆科研和教学历程，从而使本领域的研究人员和爱好者能够了解计算机视觉在中国的发展历程以及前辈们的贡献，让专委积累一些历史资料。同时，也希望通过他们的经验和视角，探讨计算机视觉及相关领域的发展现状、优势与不足，分享他们在教书育人方面的成功经验。

本次专访的是清华大学章毓晋教授。章老师属于改革开放后最早一批学习和从事计算机图像与视觉的学者，编写出版了一系列被广泛采用的教材书籍，曾任中国图象图形学学会副理事长和学术委员会主任，长期致力于倡导和推动图像工程及相关领域的学术发展。

图 1：章毓晋教授出版的图书（按出版年排列）

我是负责本次专访的主要采访人、北京工业大学贾熹滨。因疫情原因，本次采访通过邮件交流完成，相关问题由 CCF-CV 专委会的《视界专访》组提供。为能更好地帮助我们回顾本次采访，我们**采用了问答的形式来表述。**以下是章毓晋教授的简介和专访内容。

贾熹滨（采访者，后缩写为贾）：想请您和我们分享一下您的求学历程。有什么有趣的经历或难忘的故事和大家分享一下吗?

章毓晋（后缩写为章）：自己是改革开放后 1978 年 3 月入学，1982 年 1 月毕业的。毕业前考取了上海交通大学的研究生，录取后得知有部分人（大约 5%，包括自己）将被国家公派出国留学。1982 年 3 月到上海交通大学报到，得知将被派到比利时的法语区。接着，很快被先派到广州外国语学院学习法语，直到 1983 年 1 月回上海交通大学。1983 年 3 月国家联系好了比利时列日大学工学院，然后就把我们派去了。

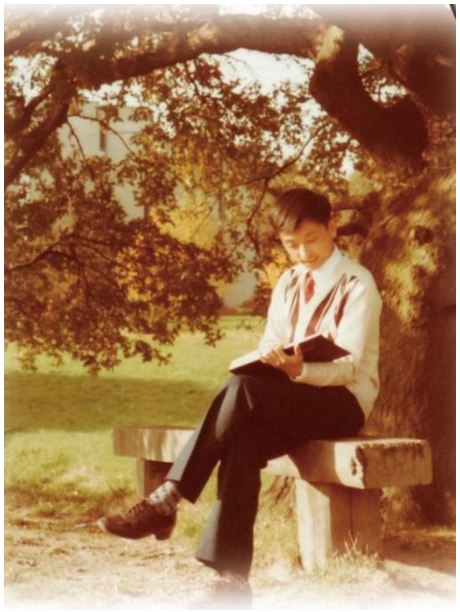


图 2：1983 年，刚到列日大学工学院校区（比利时）

到了比利时，第一件重要的事是要制订学习计划。在列日，见到了先前去留学的一些中国留学生，才知道研究生也分硕士研究生和博士研究生，如果先读硕士再读博士时间会比较长。当时改革开放还不久，国外对中国了解不多，对中国学生的水平也不清楚。一般中国学生去留学，对方院校先接收攻读硕士学位，硕士毕业后才能攻读博士学位。但当地学生就不要求这个环节，可以直接攻读博士学位。询问了学校和导师，得知如果先学习一年，考 10 门学校确定的课，如果这个“资格考试”的总成绩达到“优”，也可以直接攻读博士学位。于是自己努力学习了一年，把 10 门课都考够了成绩，从而成为了去列日大学的中国留学生中第 1 个直接攻读博士学位的学生。后来 1993 年自己回国来到清华大学时，发现学校也开始有直博生了。

还有一件当时感到反差比较大的事。出国前，国内的舆论多认为理科学生比工科学生的成绩好，有些院校会从工科学生中抽出一部分学理科，再留校。但自己出去后与当地学生一交流，发现许多学工科的学生更为自信，讲到选择工科时说出来一套一套的。反观有些学理科的学生则说起其选择时似乎有些无奈。另外，那里理科基本学制是 4 年，毕业得学士学位；而工科基本学制是 5 年，毕业得工程师学位。他们的工程师学位在当时被我们认定为国内的硕士学位。我们上学时还是国家统招统分，而在那里学生是自主择业的。由于学工科的择业机会多，收入高，所以多数工科学生比较自豪。现在国内早已是自主选科选专业，毕业后自选单位自谋职业了。另外，从创新的角度看，理科很多是要发现自然界内在的现有的规律，而工科更关注新的技术、方法、系统、产品等。工科确实有值得自豪的理由。

贾：我们了解到，您是在国外获得博士学位，也做过国外博士后和访问教授，能谈谈您认为国内和国外科研教学环境有什么异同吗？对现在国内双一流学校建设，您有什么建议吗？哪些方面国内具有的优势应该保持，哪些方面需要借鉴国外一些先进经验？

章：自己在比利时列日大学读博士 5 年多，在荷兰德尔夫特理工大学做博士后和研究人员 4 年多，在新加坡南洋理工大学做访问教授 1 年。各学校的科研教学环境各不相同，但有一些地方他们是比较相同的，且也与前些年的国内学校有些区别。那就是学校聘任教师的流程。首先，要讲一定学时的课，完成一定的教学工作量；而科研则比较自由，看教师的兴趣，相当于一个选项。我见过一个教授，每周只来学校上几次课（总学时与其他教授差不多），课外在学校里基本看不到，如果希望额外答疑需要事先约订。也见过有些教授主动申请基金项目或参与学术团体的活动，但这些都与其个人收入无关。另外，学校招了人就会分配办公室和办公设备（我读博士学位时就有独立的 10 平方米的办公室，也配备了计算机终端），与有否另外的科研经费无关。

2003 年自己在新加坡南洋理工大学学术休假时，身为教师，对他们教学管理中的一些方法当时感受比较深（但并不一定都是自己所赞同），就写了一篇“严格的章程和规范的教学”在《世界教育信息》刊物上发表，希望能对国内高校的教师同行和教学管理工作的人员有一定的启发和参考作用。其中，列出了如下几点：教学工作人员地位高、上千学生的大课、繁多的教学课时、对考题的严格审查程序、严格的考场纪律和严肃的考场气氛、对教师留任起关键作用的学生评价反馈表。这些年来，国内高校也在不断改革，好的措施和方法逐步得到普及。自己有时想，如果再去学术休假，新鲜感估计会比较小了。

贾：您是首次完整地提出了图象工程学科的概念定义，当时是什么样的契机或者出于什么样考虑，定义了这一概念？您认为这一领域的科研人员应该具备哪些方面专业素质？

章：自己是 1993 年回国来到清华大学电子工程系的（当时国内还用“图象”，约 10 年后改为“图像”。以下一般用“图像”，但原文为“图象”仍用“图象”）。因为自己的博士论文工作是有关图像分析的，所以向系里申请开一门“图像分析”课程。当时系里老师认为图像分析是图像处理的一部分，系里原已开设了“图像处理”

课程，没有同意自己新开设“图像分析”课程。自己在系里开出的第 1 门课是“计算机视觉原理”，其主要内容是有关图像理解的。

自己学习过“图像处理”相关课程，做过“图像分析”相关工作，又讲过“图像理解”相关课程，从中既感受到它们的联系，又明晰了它们的区别，所以萌生了将它们分层次联系在一起，以全面覆盖图像技术领域的念头。事实上，随着图像技术研究的不断发展，单用“图像处理”已经很难覆盖相关领域的研究内容了；而随着图像技术应用领域的不断扩展，其与社会生活发展结合得也越来越紧密，仅考虑“图像处理”也不够全面。所以，需要有一个比“图像处理”更（概念层次）“高”和更（覆盖范围）“大”的名称来统领更广泛的领域。由于自己正好在电子工程系，所以就决定用“图像工程”来统领它们（工程是指将自然科学的原理应用到工业部门而形成的各学科的总称）。通过综合调研，确定了图像工程的内含外延，提出“图像工程是一个系统地研究各种图像理论，开发各种图像技术，以及研制和使用各种图像设备（应用于广泛领域）的综合学科，主要可分成紧密联系又有区别的 3 个层次：图像处理、图像分析和图像理解，还包括对它们的工程应用”。通过对三个层次在操作对象和语义层次上的特点以及在数据量和抽象性方面的区别的讨论，构建了图像工程研究的三层次模型。

第一次使用“图像工程”这个名词是 1995 年。自己到电子工程系后，系里希望自己开门英文课。经过全面准备，自己从 1995 年开出了“Special Topics in Image Engineering”。

在 1996 年，自己在一个国际会议上的报告中正式提出了图像工程学科的概念和三层次模型。同年在新创刊的“中国图象图形学报”开始撰写“图像工程”综述系列。这个综述系列至今已有 27 年，目前还在进行中。这 27 年的综述均发表在“中国图象图形学报”每年 5 月份那期上。每篇综述都对上一年发表在国内 15 种重要刊物上的图像工程相关文献(至今已达 17535 篇)进行选取、统计和分析。随着国内外学术交流的广泛开展和国内研究水平的不断提高，统计的信息也越来越反映了世界范围内相关科研工作的内容和趋势。同时，自己还积极参加了“中国图象图形学报”的栏目建设。截止目前，自己建议的三个栏目：图像处理和编码，图像分析和识别，图像理解和计算机视觉，仍是学报最主要的栏目。以对 2015 年学报的统计为例，在全年学报曾出现的 10 个栏目中，只有“图像处理和编码”以及“图像分析和识别”两个栏目是每期都有；这两个栏目下的文章数量，比其余八个栏目下的文章数量的总和还要多。因为科技文献的发表是科研人员研究成果的一种体现，其内容也反映了当时研究领域的范围和特点，反映了理论研究的深入、工程技术的发展、应用领域的拓展，所以从中可看出该领域的总体研究情况。

对学科的基本原理，还需要有教材来阐述，有课程来讲授。从 1997 年开始，自己为系里其他教师编写了“图像处理和分析”讲义，用于“数字图象处理基础”课程。从 1999 年开始，自己陆续开始编写和出版图像工程系列教材，第 1 版包括“图象工程（上册）：图象处理和分析”、“图象工程（下册）：图象理解与计算机视觉”和“图象工程（附册）：教学参考及习题解答”。其中，上册继续用于“数字图象处理基础”课程，还用于为外校开的“数字图象处理”课程；下册则继续用于自己原为系里学生开的“计算机视觉原理”课程（那时已改名为“图象理解与计算机视觉”）。

顺便说一下，“计算机视觉原理”课程从 1996 年开始改名为“计算机视觉与图象理解”，从 2000 年开始改名为“图象理解与计算机视觉”，从 2006 年开始改名为“图像理解”。2003 年，自己在新加坡南洋理工大学学术休假一年。当时，那边希望自己开一门课，自己就提出并开出了“Advanced Topics in Image Analysis”课。有了这个基础，2004 年学术休假回来后，自己再次提出开“图像分析”课，就马上得到了批准。同时，由于原讲“数字图象处理基础”的老师退休，自己接手了该课，并从 2006 年开始改名为“图像处理”。

后来，这 3 门课，即“图像处理”、“图像分析”和“图像理解”，就一直讲了下来。所用的教材依次是自己编写的图像工程第 2 版、第 3 版和第 4 版。每版都包括 3 册：“图像工程（上册）：图像处理”、“图像工程（中册）：图像分析”和“图像工程（下册）：图像理解”。另外，自己 2009 年编写了“Image Engineering: Processing, Analysis, and Understanding”。2017 年还编写了“Image Engineering, Vol.1: Image Processing”，“Image Engineering, Vol.2: Image Analysis”和“Image Engineering, Vol.3: Image Understanding”。

这里顺带简单谈一下自己对图像工程与计算机视觉关系的看法。首先，它们是密切相关的。图像是表达视觉信息的一种物理形式，对图像的操作也需要借助计算机来进行。计算机视觉作为一门学科，与许多以图像作为主要研究对象的学科，特别是图像处理、图像分析、图像理解有着非常密切的联系和不同程度的交叉。计算机视觉主要强调用计算机实现人的视觉功能，这中间实际上需要用到图像工程三个层次的许多技术，虽然目前的研究内容侧重于高层视觉且主要与图像理解相结合。历史上，也有人称计算机视觉为图像理解。事实上，图像理解和计算机视觉这两个名词也常混合使用。从本质上讲，它们互相联系，在很多情况下其内容交叉重合，在概念上或实用中并没有绝对的界限。在许多场合和情况下，它们虽各有侧重但常常是互为补充的，所以将它们看作是专业和/或背景不同的人习惯使用的不同术语更为恰当。一般来说，对这些相关内容，计算机系有计算机背景的人更多称之为计算机视觉，电子工程系有信号处理背景的人更多称之为图像处理，而自动化系有模式识别背景的人常称之为图像模式识别。

除了教材，为对图像工程学科提供支持，自己 2009 年写出了《英汉图像工程辞典》，当时只有 2000 个词条。2015 年出版的第二版包括了五千个词条，而 2021 年出版的第三版则包括了上万个词条。另外，2021 年 Handbook of Image Engineering 由 Springer Nature 出版，除出版纸质书外，出版社还出版了电子版（包括 PDF 和 EPUB 两种），目前下载已达 6 万多次。

从学习图像工程课程的角度来说（如教材中所写），有 3 个方面的基础知识是比较重要的。

(1) 数学：首先值得指出的是线性代数和矩阵理论，因为图像可表示为点阵，需借助矩阵表达解释各种加工运算过程；另外，有关统计学、概率论和随机建模的知识也很有用。

(2) 计算机科学：计算机视觉要用计算机完成视觉任务，所以对计算机软件技术的掌握，对计算机结构体系的理解，以及对计算机编程方法的应用都非常重要。

(3) 电子学：一方面采集图像的照相机和采集视频的摄像机都是电子器件，要想快速对图像进行加工，还需要使用一定的电子设备；另一方面，信号处理是图像处理的直接基础。

图 3：（从左至右）：李卓，沈渊，李国林，邓北星，黄翌东，章毓晋，何芸，李东梅，黄永峰，孙长征
2020 年 1 月 2 日，清华大学电子工程系（北京）

贾：数字图像处理研究有五六十年历史。您认为哪些问题已经被很好地解决了？哪些还没有？原因是什么？

章：前面提到的综述系列里主要将相关文献划分为“图像处理”、“图像分析”、“图像理解”和“技术应用”四大类。从图像工程研究的角度看，所提问题主要对应前三个大类。这三个大类在这 27 年的文献数量统计情况可见下图。图中还用粗直线给出了三个层次文献数量的（线性）趋势线。从总体数量来看，“图像处理”和

“图像分析”文献数量较多，“图像理解”文献数量较少。从总体变化趋势来看，“图像处理”是向下的，“图像分析”是向上的，“图像理解”也是向上的。研究文献的数量在一定程度上反映了研究者对不同问题的关注程度以及在相关方向所取得的成果。据此图分析，“图像处理”方面的问题已被解决掉的问题相对比较多，新的研究相对比较少；“图像分析”方面的问题近年正得到广泛关注和研究，有不少问题已开始有了较满意的解决方案，但还有更多问题在研究中；而“图像理解”方面的研究开展相对比较少，未解决问题还有不少，但相应研究正在逐步加强。对此更具体的分析和讨论还可参见上述综述系列。

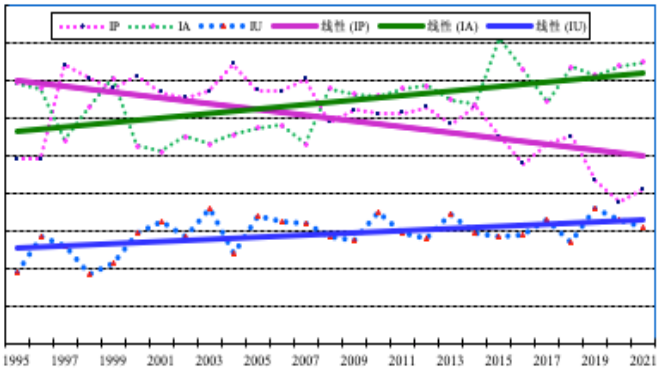


图 4：从图像工程研究的角度看，三个大类 27 年的文献数量统计情况

从计算机视觉的角度来说，计算机视觉要用计算机来实现人类视觉功能。视觉功能包括（较低层次的）视感觉功能和（较高层次的）视知觉功能。限于目前技术水平，用计算机实现视感觉功能已有相对较多的成果，而用计算机实现视知觉功能还有许多工作要做。

贾：您在清华大学，先后开出并讲授 10 多门本科生和研究生课程，出版了一系列优秀的图像、视觉领域教材，请您谈一谈您在教学方面的心得体会。能否谈谈您是如何平衡教学和科研的工作，有什么经历和我们分享吗？对目前教学科研一线的青年教师有什么建议吗？

章：教师要教书育人，所以开课讲课应是很多教师的一个基本任务。如前所述，国外许多学校对教师的教学工作量有明确的要求，对科研的要求则不那么硬性（当然，科研成果突出对教师的学术影响力和职称提升都是必要的，但好像还不是聘任的充分条件）。近年清华大学对教研系列教师也有了明确要求（80 学时/每年），已成为各级聘任的必要条件。

图像技术作为近年发展较快的工科专业，应该不断有新的课程开出来。除了图像工程三个层次的“图像处理”（对不同学习对象内容也不同）、“图像分析”和“图像理解”课，以及“计算机视觉”和“图像工程专题”外，自己还结合科研工作开设了“基于内容的视觉信息检索”，组织所里老师开了“图像新技术”。除了开新课，已开课其内容也应结合科研、不断更新。如自己在参与开设的“电子学与通信学科前沿”中，从 1998 年到 1999 年讲的是图像工程，2000 年到 2004 年讲的主题是数字水印，2005 年到 2008 年讲的主题是表情分类，2009 年到 2010 年讲的主题是人脸识别，2011 年到 2013 年讲的主题是人脸分析，2014 年后讲的主题是时空行为理解。同时，每年在相同主题下也调整介绍一些新的技术。

开课讲课就涉及到教材建设的问题。教材应注意其理论性、实用性、系统性和实时性。合适的教材对课程的质量起重要的作用。作为专业性的课程，讲课人自编的教材用起来会更得心应手。

人们常说“教学相长”，其中一个意思应是教师从与学生的交流中也有收获。事实上，自己也从讲课等教学活动中受益匪浅。自己写的教材能被许多人使用，其中也与自己在讲课以及与同事、学生的交流中了解了一些大家关注的问题（以更新内容），大家容易误解的问题（以改进描述），大家不太好理解的问题（以加强解释），从而不断改进有关。上个世纪及本世纪初，国内出版的图书上一般没有作者联系方式。自己在写前言时把自己的联系方式附在后面，编辑还提醒说如果有人来联系作者可能会很忙。结果确实是有很多人来电来邮件，提出意见、建议、问题等，也占用了一些时间。但自己收到了这么多反馈，对讲好自己的课以及对改进教材的新版内容、表述方式、结构形式也很有启发和帮助。

教学和科研是互相促进的。教学强调系统性，对一个专业的系统把握肯定对开展相关专业的科研有帮助。科研强调创新性，通过科研肯定会对理论有更深的理解，将科研成果结合进教学也会提升教学效果。我们是专业教师，这种互相促进是很有益的，自己也乐于结合进行。在清华大学工作的 20 多年间，自己讲课学时平均约为每年 110 学时。自己独立讲授过的 9 门课中，有 8 门是新开出来的，另一门也是先写出教材由其他教师讲到退休才接过来的。自己参与讲授的 5 门课中，有 3 门是自己组织开出来的。事实上，对教师来说，教学不是负担，而是一个很好的机遇。

顺便说一下关于承担科研项目的问题。自己以为最好还是要承担偏重研究的项目（如国家自然科学基金等）。关于承担横向项目，自己比较偏向于和国外公司的合作，因为相对来说合作的研究性质比较强，不要求学校教师完整地做一个类似于产品的结果出来，而且研究工作完成之后常常还可以发表成果。近年来，随着国家对研究的投入越来越多，学校对新入职教师的资助力度越来越大，教师开展科研工作对项目的依赖比前些年已少了。

贾：您在南洋理工大学，也开出并讲授过研究生课程：“现代图象分析（英语）”，经常听到大家说国内学生在课堂上参与度不如国外学生，能谈谈您的感受吗？分享些您和学生相处的故事吗？您认为一个优秀的学生应该具备什么样的素质?可以对现在学生包括不在清华等一流学校的普通学校学生在图像工程领域等专业学习提升科研素养方面给一些建议吗？

章：我经常说，学生都是同样聪明的，但有可能聪明在不同的地方。从课程学习来说，课程首先要看内容，如果学生感兴趣内容，比较重视课程，就会有较高的参与度。其次要有一定的授课方式，能够调动学生的参与度。但这与学生基础、课程特点等都有关，不能一概而论。近年来，国内高校学生选课的自主性有所提高，自己感觉学生在课堂上的参与度也更高了。

在我们上学和出国留学那段期间，相对来说，国外学生选课自由度较大。对感兴趣的课程，有一部分国外学生不仅课堂上参与度比较大，课外作业、实验等的参与度也比较大。记得当时先去的中国留学生说，一般在一个班里，有少数外国学生很突出，你要想成绩比他们好不容易；但大多数也比较一般，你努力超过他们也是可做到的。特别是总能遇到一些外国学生比较早就自己选专业、选课程，一学起来觉得不是自己所想象或期望的，下一年换个专业退一班又重新开始学。

前些年，清华本科生毕业后出国深造的比较多，外校招来的硕士生和博士生也比较多。从多年的培养效果来看，两种来源的学生（包括自己带的和自己周围教师所带的）均有学习和科研都很努力，课程成绩好，科研成果突出的；但也都有忙于其他事物，课程成绩不理想，科研工作深入不下去的。当然，好的知识基础、好的学校、好的老师和同学、好的环境和氛围对提升科研素养都有好的影响，但个人的认识、兴趣、期望目标、专注程度和努力程度往往起到的作用更大。

贾：现在深度学习在包括图像分析、理解等计算机视觉领域越来越占据重要的地位，您是怎样看待这一现象？对于现在的研究人员，在这样一个环境下，应如何看待经典图像领域理论方法和应用呢？您对从事该领域科研的年轻人有什么建议？

章：人工智能本身就是一个很大的学科，深度学习是当前发展比较迅速的一个领域，但自身也还存在一定的问题尚未很好地得到解决，还需要相关研究人员进一步的工作。

对图像工程或计算机视觉领域来说，深度学习是一个有用的、推动图像技术发展的重要工具。历史上，像小波理论、人工智能、模糊理论、神经网络、遗传算法、机器学习等理论方法提出后，基于它们的各种图像（处理、分析、理解）技术都曾经更新过，也确实得到了很多应用。应该说，深度学习看起来会比先前的许多理论方法更强一些，但要完成涉及图像的各种任务，还是需

要先把握住图像和图像技术的自身特点和规律，才能更好地构建深度学习的模型和网络，取得好的效果。

贾：看到您承担了多届中国图象图形学会的旗舰会议，包括从第一届开始的国际图象图形学术会议以及多次全国图象图形学术会议的程序委员会主席，可以请您谈谈开始组织这些会议的初衷吗？能分享一下初期举办的一些故事吗？

章：从 2000 年开始，中国图象图形学学会开始主办国际图象图形学学术会议 (ICIG)。在前八届会议中，自己担任了七届的程序委员会主席（另一届自己在外学术休假）。另外，从 1982 年开始，中国图象图形学学会开始主办全国图象图形学学术会议（NCIG）。自己在 2005 年到 2018 年的连续八届会议上也都担任了程序委员会主席。

从左至右：游志胜，章毓晋，潘云鹤，谭铁牛，韦穗
图 5：2005 年 10 月 12 日，NCIG2005 开幕式 (北京)

首届国际图象图形学学术会议是 2000 年在天津召开的。当时天津图象图形学学会获得了承办中国图象图形学会 2000 年第十届全国图象图形学学术会议（NCIG2000）的资格，然后提出申请将该次会议改为 2000 年第一届国际图象图形学学术会议(ICIG2000)。在得到中国图象图形学会批准后，天津学会开始筹备，自己被提名担任了此次会议的程序委员会主席。主要工作一是要组织邀请一些国际专家担任大会主席、程序委员会主席，做大会特邀报告等；二是要组织程序委员会，负责审稿，制订会议程序，出版论文集；三是要起草征文通知，并向国内外散发。当时国内举办国际会议不多，有些人对相关情况了解不多。记得有一次筹备会上，讨论参会人员胸牌的格式，有人提出要将境外参会人员的护照号码也印上去。自己赶快进行了解释说明，使胸牌得以规范印制。自己还借出国参加国际会议的机会，携带了很多份征文通知。除广泛散发以外，还利用会间休息，共进午餐，参加晚宴等机会，一对一地向与会境外学者介绍中国情况、学会情况、会议情况。最后，经过学会的组织和宣传还有参与者的努力，虽然该次会议是该系列从无到有的第一届，但此届会议至今仍是该系列会议中境外参会人员最多且所占比例最大的一届。

从左至右：章毓晋，李卫平，徐冠华
图 6：2011 年 8 月 12 日，ICIG2011 开幕式（合肥）

图 7：2020 年 12 月 5 日，中国图象图形学会成立 30 周年庆祝大会（北京）

贾：目前学界的工作成果越来越丰富，也吸引了越来越多的研究者参与，但是要做出真正有价值的工作是很难

的，您认为要做出有价值、影响力的工作，需要做出哪些努力呢？

章：马克思曾说过："在科学上面是没有平坦的大路可走的，只有那在崎岖小路上攀登不畏劳苦的人，才有希望到达光辉的顶点。"

鲁迅曾说过："伟大的成绩和辛勤劳动是成正比例的，有一分劳动，就有一分收获，日积月累，从少到多，奇迹就可以创造出来。"

贾：请您对 CCF-CV 专委简报的读者寄语。

章：随着社会发展和技术进步，计算机学科特别是计算机视觉专业迎来了一个前所未有的机会。祝各位同仁积极努力，抓住机遇，不负使命，投入到相关的研究和应用中，为国家、为社会、为人民贡献自己的力量，这是最幸福的！

责任编委 贾熹滨 张军平 明悦

章毓晋

1989 年获比利时列日大学应用科学博士学位，1989 年至 1993 年先后为荷兰德尔夫特大学博士后及研究人员，1993 年到中国北京清华大学电子工程系工作至今，1997 年起聘为教授，1998 年起成为博士生导师，2014 年起聘为教研系列长聘教授。2003 年学术休假期间曾为新加坡南洋理工大学访问教授。在清华大学，先后开出并讲授过 10 多门本科生和研究生课程。在南洋理工大学，开出并讲授过研究生课程：现代图像分析（英语）。主要科学研究领域为其积极倡导的图像工程（图像处理、图像分析、图像理解及其技术应用）和相关学科。已在国内外出版图书 50 多本，发表图像工程研究文章 500 多篇。曾任中国图象图形学学会副理事长和学术委员会主任，第二十四届国际图象处理学术会议（ICIP2017）程序委员会主席。现为中国图象图形学学会会士和名誉监事长；国际光学工程学会（SPIE）会士。邮箱：zhang-yj@tsinghua.edu.cn ， 主 页 ： http://oa.ee.tsinghua.edu.cn/zhangyujin ； http://web.ee.tsinghua.edu.cn/zhangyujin

心底无私视界宽∞阮秋琦教授专访

自 50 年代以来，我国在计算机视觉领域展开了相关的科研工作。而今，我国已经拥有了一支庞大的、在这一领域辛勤耕耘且能与世界一流水平并驾齐驱的科研队伍。在这一过程中，有一批见证了视觉领域的发展，为我国计算机视觉领域的奠基做出了重大贡献的先驱。

《CCF-CV 专委简报》视界专访栏目希望通过对计算机视觉研究历史、进展的见证者作一个系列专访，以帮助从事计算机视觉及相关领域的科研工作者或爱好者，全方面地了解 50 年代以来信息技术、信号处理技术以及计算机视觉相关的一些历史发展及进步，也希望能帮助我们在见证这段历史的同时，展望计算机视觉领域的未来。

负责本次专访的主要采访人是北京邮电大学明悦。因疫情原因，本次采访通过邮件交流完成，相关问题由视界专访栏目组提供。为能更好地帮助我们回顾本次采访，我们**采用了问答的形式表述**。以下是阮秋琦教授的专访内容和简介。

问：想请您和我们分享一下您的求学历程。有什么有趣的经历或难忘的故事和大家分享一下吗？

阮教授：我的求学经历很简单，从小学、中学、大学应该说都很顺利。小学到中学，中学到高中是保送的。1964 年考大学也比较顺利。那时与现在大不一样，当时同龄人平均每百人只有 1 人能上大学，可见竞争之激烈。我的中学老师书教得非常好，但都没有上过大学。由于家在小县城，报考志愿基本没有概念。想起来报考的经历

图 1 阮秋琦教授

一是凭兴趣，从玩矿石收音机开始对无线电和自动控制很着迷；二是人们普遍崇拜铁路行业；三是很向往北京。因此报考了北京铁道学院，幸运的是顺利地考上了。在大学我们扎扎实实地学习了两年基础课，后来由于众所周知的原因，中断了学业，直至在复课闹革命的口号下，又学了两年专业基础课和专业课，回想起来还是打下了坚实的知识基础。1969 年毕业后留校工作，当时学校近 10 年没有招生，我就与大多数老师一样分配到电信系信号工厂做研发工作。当时研究电缆串音测试仪和电平测试仪，专业知识和实践能力大有提高。跟随蒋焕文老师也学到很多东西，至今想起来受益匪浅。后来工农兵学员开始进校，在工厂的老师陆续回归教学岗位，我就到恩师袁保宗老师的研究室工作，直至现在。后来 78 年我国恢复了研究生制度，在袁老师的支持下，我又考取了首届研究生，有了进一步深造的机会。当时我校是

图 2 阮秋琦在科研课题中与导师交流

首批能招收研究生的院校之一，我们这个学科在全国招收了 12 名研究生。经过三年的刻苦学习和严格的训练，于 1981 年毕业，并获得硕士学位。我们当时在京院校的毕业研究生学位授予仪式是在人民大会堂举行的，国务院发布了公报，刊登了所有获得学位人员名单。我至今还保留着，当时的情景至今难忘。大家都知道“兴趣是最好的老师”，我报考大学和选择专业就是凭着兴趣，同时也有强烈的求知欲望。在高校停止招生和没有教学科研任务的一段时间内，我有幸能留校从事科研工作，是较好的动手能力使我快速进入角色，承担和完成了比较重要的工作。在业余时间，我给很多老师组装了晶体管收音机，在复课期间自己动手设计和组装了第一块万用表，由于比较喜欢音乐，自己组装了高音质电子管音响，同时组装了我校第一台电视机（当时国内没有卖电视机的）。每当晚饭后，老师、朋友、邻居都来我家看小电视，也是一段很难忘的趣事。在科研的基础上，我还负责仪器仪表的维护维修工作，那时几乎所有的仪器仪表我都能熟练地使用和维修，这需要涉及广泛的理论知识和娴熟的动手能力，但也无形中奠定了今后在教学、科研和深造中能得心应手的基础。

我在教学和培养学生方面十分注重学生三个能力的培养，即：自学能力、动手能力、外语能力。我在国外经历的一件小事进一步印证了三种能力的重要作用。在 1985 年我去美国时，这一基本功使我受益匪浅。那会，我在“Computer Vision & AI “ 实验室工作，刚一去导师就要我修复一台 PDP-11 计算机。在那个时候，

图 3 阮秋琦教授指导研究生

国外的计算机也是很贵重的实验设备，而我国计算机还很不普遍，读研究生时只用过国产的 DJS121 机，非常原始，输入要穿纸带，PDP-11 计算机根本没见过。同实验室的美国同事告诫我，最好不要接这一棘手的难题，这之前曾有五位博士都没有解决这一问题，但我想我不能进实验室伊始就给人留下一个能力有限的印象，在美国不可轻易说“不”，至少也要试一试。我相信再复杂的机器也是由最基本的硬件电路和基本单元组成的，根据我自己的经验和能力，可以一试。经过仔细分析和严格的测试，终于发现一块集成电路不正常，导致磁盘不工作，使整个机器无法工作，更换之后，机器马上复活了，可以说一炮打响。我是凭借着自己坚实的基础理论和突出的动手能力，在一周之内彻底解决了这一问题，取得了导师的信任。自此以后，每当在走廊遇见美国同事时，他们都会友善而礼貌地说“you are computer expert”。这次使我再一次尝到了掌握坚实的基础理论和突出的动手能力的甜头。虚心好学使我受益多多，一个人任何技能都值得学习，而且一生中总有用武之地。

问：我们了解到，您曾在 90 年代多次赴国外进行访问研究，能谈谈您认为国内和国外科研教学环境有什么异同么？国外进行的研究工作对您回国后继续进行的研究有哪些益处？对现在国内双一流学科建设，您有什么建议？哪些方面国内具有的优势应该保持，哪些方面需要借鉴国外一些先进经验？

图 4 阮秋琦教授访问俄罗斯

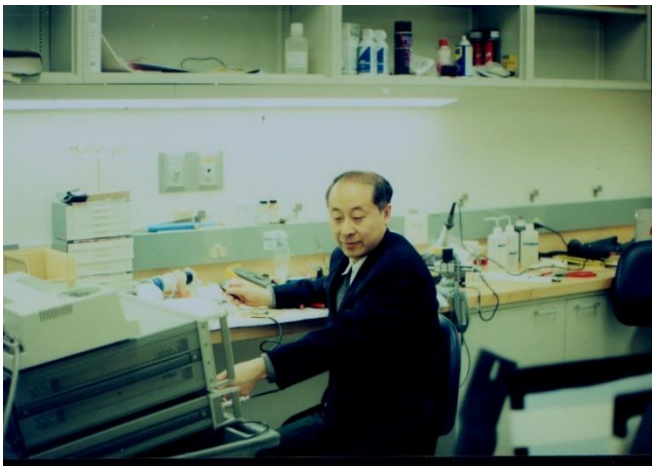


图 5 阮秋琦教授在美国匹兹堡大学作客座研究

阮教授：我是 1985 年底公派出国进修的，由于需要，在美国学习工作了近四年时间，参与了一些当时匹兹堡大学和辛辛那提大学的科研工作，后来又有两次去该校客座研究的机会。总体看来国外和国内的人才培养各有千秋，要说有差别，第一、是国外的学术环境比较宽松、自由，特别是大学教师，考核比较宽松；第二、院校极其重视教师队伍建设，如有一年 CMU 得到一笔不小数目的私人捐款，这些钱没有用于学校的基础设施建设，90%的钱用来招聘经济学领域的知名学者，使得该校的学科有了长足的进步；第三、教学中鼓励学生积极思考，课堂教学比较活跃，学生提问非常积极，这一点我们的学生还有差距；第四、国外大学的学风值得我们学习，校园服务周到而细致，读书的习惯普遍好于我们现在的大学；我们的大学特别是深度阅读的习惯还要进一步加强。如哈佛大学有 100 多座图书馆，匹兹堡大学和 CMU 也有很多图书馆，几乎每个系都有自己的图书馆，图书借阅十分方便，服务也相当周到。有一次我到图书馆借一本书，管理人员告诉我，馆里没有这本书，要我等一个星期，他从宾夕法尼亚大学给我调来。一周后果然通知我去取书。CMU 也是，在 CMU 的图书馆读书，只要把书名写在纸条上放在桌子上，管理人员就会帮你找到书，并放在你的面前；读完后，把书放在桌上走人就可以了。图书馆内极其安静、宽敞，在那里看书简直就是享受。

国内大学要想建成一流大学，达到双一流建设目标，我觉得至少有三件事要做：第一就是师资队伍建设，这在中国早有定论，如蔡元培校长的名言“所谓大学者，非谓有大楼之谓也，有大师之谓也”。队伍建设是双一流建设的重中之重，队伍建设要坚持老中青三结合，引进与培育相结合的方针，充分发挥学术造诣深厚，德艺双馨的老教授的传承作用，因为没有传承就没有发展；第二就是抓学科建设，这是人才培养之根本；第三就是加强校园文化建设，校园文化涉及方方面面，没有一流的校园文化不可能建成一流的学校，也就不可能培养出一流的人才。在我国，西南联大就是一个突出的例子，在那战火纷飞的年代里，高校被迫西迁，成立的西南联大，办学条件极其艰苦和简陋，却培养出一大批蜚声国内外的杰出人才，不能不说是得益于西南联大的优良的校园文化。我想这三件事如果有所突破才能达到双一流的建设目标。

问：您长期从事信号与信息处理领域的教学和科研工作。承担国家及省部级项目多项，取得了很多的科研成果，能谈谈科研项目和研究成果取得、人才培养、学科建设等方面的联系吗？

阮教授：研究生毕业后我长期从事信号与信息处理方面的教学和科研工作，承担国家、省部级项目 80 余项，也取得了一些成果。总体看，教学、科研、社会服务是高校教授的基本任务，这些工作是休戚相关，互为补充的关系。目前多数学校在教师晋升都设了教学科研型、

图 6 阮秋琦教授（后排左一）在超级智能视听信息处理系统研究课题汇报会现场

教学型、科研型的不同岗位，我个人认为这是队伍建设的权宜之计，长远看还是要鼓励教学科研并重，这样才能建成一流的教师队伍。例如：大学本科阶段，课堂上的理论学习占较大的比重。它为学生奠定了研究创新必不可少的知识基础，理论教学也是“创新性”思维训练的重要途径。任何一门科学或课程都有自己的科学体系及独特的魅力，这些体系的形成本身就是经过科学工作者长期研究总结的结果，在课堂教学中大力突出具有“创新性”思维训练价值的内容，把它们作为重要的知识点，强调创新性思想的意义及价值作为教学的主线，着力培养学生的“创新性”思维习惯。

我们以“数字图像处理”这门课作为研究载体进行了“研究性”教学试点，图像是视觉信息，形象化的处理效果对学习者会产生强烈的视觉冲击。同时，在学习开始就在绪论一章给出大量实例，从而激发学生们的学习兴趣，这正是本门科学的独特魅力所在。利用本门科学的魅力激发学生的学习兴趣是“研究性”教学法的第一步。在激发学生学习兴趣的同时，提炼出“研究性”思维的切入点是我们教学研究的重点之一。数字图像处理中涉及大量的数学问题，我们不是枯燥地讲解数学理论，而是从工程应用的角度使这一数学工具回归工程，既强调数学的严密性，培养学生严密的逻辑思维能力，又使学生了解工程运用的规律，从而使学生得到“创新性”思维训练。如果没有深厚的科研功底和经验，在教学中不可能有如此生动的表述，也就不可能达到我们工科院校的培养目标。所以科研是教学素材的积累过程，教学是理论成果的总结与升华过程，他们之间是互补的。

问：您出版书籍和译著多部，我们中的很多人最早是看了您的《数字图像处理》，从而走进计算机视觉这个研究领域的，您能谈一谈如何撰写一部优秀的教材和专著么？在撰写这些书籍和翻译国外学者的著作过程中，您有哪些收获是希望和青年学者和学生分享的么？

阮教授：在我几十年教学科研中，比较注重教科书的编著工作。最早在 1975 年就参与了“低频电子电路”教材的编著，这本书是蒋焕文老师主编，我们参与，这在当时是很有影响的一本教科书。之所以有影响是由于这是理论与实践相结合的一本教材，有特色，书中的每一种电路都有理论分析和实验总结，每种典型电路都是我们亲自搭建电路、亲自实验得出的结果，既有严密的理论分析，又有实践经验总结，受到读者和学生的欢迎。这本书被匹兹堡大学的东亚图书馆收藏了，是该图书馆看到的唯一一本中国的教材。当时处在文化大革命期间，没有作者署名，只是署名北方交通大学编写组。我编著的第一本“数字图像处理基础”一书是 1982 年，当时我校是第一批开设“数字图像处理”课程的院校之一。当时教材缺乏，所以 1982 年我编写了一套讲义，共三本，经 3 年的教学和科研实践，于 1985 年正式列入出版计划，由中国铁道出版社出版。这是我们国内比较早的一本图像处理的教材。后来进一步总结提升，于 2000 年编著了“数字图像处理学”一书，如今已是第 4 版。我个人认为教材是一种特殊的著作，教材建设是教学活动的基础工作和教学改革成果的结晶，体现了人才培养计划中知识结构设计的精髓。一部好的教材应具有如下特点，即系统性、科学性、完整性、先进性、实践性。教材的受众广泛，从初学者到各个领域有需要的专业人士都可以阅读，而且能读懂，一般只有成熟的科学和技术才能进教材。而专著有自己的特点，专著一般受众范围较小，通俗地说是给小同行阅读的书籍。一般初学者

很难读懂，或者要参考大量的相关书籍才能读懂。我编著的教材就是本着这样的宗旨，因此才能比较受欢迎，被评为“十一五”“十二五”规划教材，也获得优秀教材、国家级精品教材的奖项。“数字图像处理学”被评为国家级精品课，具不完全统计，已有近 70 所院校用作教材和教学参考书，总计发行 10 余万册。

问：您翻译的书中，图像处理的很多问题，都是以图像处理特有的方法来实现的。而现在的图像处理，在很多应用或问题的处理，都已经被更为通用的深度学习模型所取代。您如何看待这一现象？另外，未来是否还需要图像处理领域特有的理论或算法?

阮教授：数字图像处理起源于二十世纪二十年代，由于遥感、医学等领域的应用，使图像处理技术逐步受到关注并得到相应的发展。1964 年美国的“喷气推进实验室”处理了由太空船“徘徊者七号”发回的月球照片，这标志着第三代计算机问世后数字图像处理开始得到普遍应用。由于 CT 的发明、应用及获得倍受科技界瞩目的诺贝尔奖，图像处理技术大放异彩。其后，数字图像处理技术发展迅速，目前已成为工程学、计算机科学、信息科学、统计学、物理、化学、生物学、医学甚至社会科学等领域中各学科之间学习和研究的对象。

随着信息高速公路、数字地球、物联网等概念的提出以及 Internet 的广泛应用，信息传输中的非话业务也随之急剧增长。其中，图像信息以其信息量大，传输速度快，作用距离远等一系列优点使其成为人类获取信息的重要来源及利用信息的重要手段。图像处理又是与国计民生紧密相联的一门应用科学，它已给人类带来了巨大的经济和社会效益，它的发展及应用与我国的现代化建设联系之密切、影响之深远是不可估量的。但图像处理仍然是十分活跃的一门科学，在理论体系的建设上仍然有相当多的工作要做。我认为以下四点仍然是图像处理科学进一步发展的主要任务，（1）在进一步提高精度的同时着重解决处理速度问题。（2）加强软件研究、开发新的处理方法，特别要注意移植和借鉴其他学科的技术和研究成果，创造新的处理方法。（3）加强边缘学科

图 7 美国匹兹堡大学 CC Li 教授访问信息所（左：阮秋琦教授，右：CC Li 教授）

的研究工作，促进图像处理技术的发展。（4）加强理论研究，逐步形成图像处理科学自身的理论体系。

冈萨雷斯的“数字图像处理”一书是我建议电子工业出版社引进的。当时电子工业出版社成立了一个编委会，由清华大学吴佑寿院士为组长，专门负责国外教材引进推荐工作，我是成员之一。我推荐了冈萨雷斯的这本书。在美国期间我通读了他的全部著作，也读了如罗森菲尔德、因奈斯特.豪和普拉特的“数字图像处理”，我觉得冈萨雷斯的书很有特点，推荐给出版社。后出版社委托我翻译成中文，这本翻译本被评为全国最畅销书。

深度学习作为图像处理科学与技术的一种方法在特定任务域中确实取得了令人瞩目的成果，但我个人认为这并不是说其他经典的方法就可以忽略了，相反以上四点发展方向仍然是从事图像处理科学研究的主要任务。更何况深度学习自身的诸多问题也没有得到很好的解决，其实深度学习的理论基础还不够坚实。在图像处理工程中的 5 大研究领域和 8 项处理内容中，深度学习还是整个图像处理中的一部分，我相信传统技术的研究随着新的数学理论和工具的应用还会有更多有效的方法问世。

问：我们了解到，您除了取得了很多优秀的科研成果外，还获得了国家级教学名师、宝钢优秀教师特等奖等多项教学领域的奖励和荣誉，请您谈一谈您在教学方面的心得体会。能否谈谈您是如何平衡教学和科研工作，有什

图 8 阮秋琦教授（第一排右二）参加博士答辩会

么经验和我们分享吗？对目前教学科研一线的青年教师有什么建议么？

阮教授：高校教师的主要任务是培养人才，教学仍然是教师的重要任务。我在近 50 年的教学工作中做了一些有益的探索工作，获得了国家、学校及同学们的肯定。我认为提高办学和人才培养质量除了要有良好的学风、先进的教学理念、恰当的培养计划、全面的课程设置及充分的教学条件之外，重要环节之一就是要有高水平的教师、高水平的课程、高水平的教材、高水平的授课质量。要讲好一堂课，要有好多积累，众多条件，长期锻炼，细心研究。一位好的教师是经过多年的磨练而自然形成的，是经过多年的潜心修炼而水到渠成的结果。就教学来说，我个人以为一位好老师还是要注意几个基本要领。

1. 心存敬畏，方能精益求精。作为大学教师，教书是他的主业，传道、授业、解惑是教师的天职。因此，课堂是传授知识的圣地，容不得半点懈怠与敷衍。对课堂教学没有敬畏之感，就难当此重任；

2. 厚积薄发，方能游刃有余。这就要求教师具有坚实宽广的基础理论和系统深入的专门知识。也就是要有深厚的学术功底。这是不断学习、长期积累的结果，经过长期有准备的积累，才能上好一堂课；

3. 精雕细刻，方能深入浅出。教学是传道、授业、解惑的过程，每一个名词、术语和基本原理都要科学、准确地加以诠释，来不得半点模棱两可。把复杂的课程内容通俗易懂地讲授给学生就必须对教学内容精雕细刻；

4. 科研引领，方能激发兴趣。一名优秀的教师，必须是长期从事本门科学的科研及教学的专家，而且对相关学科的知识也颇有造诣，这样才能深刻理解本门科学的精髓，从而以科研案例激发学生的学习兴趣；

5. 手段多样，方能形象生动。任何一门科学或一门课程都有其独特的魅力。我们教师的任务就是通过多年的教学和科研经验把这一魅力的内涵挖掘出来，以通俗生动的手段展现给学生，这是教学的基本要求。

6. 旁征博引，方能记忆深刻。已学知识的经常引用，相关知识的联想思维是教学中加深记忆的重要方法。在课堂讲授中要重视知识冗余的运用。

总之，一位教师要时刻遵循这六个要点，加以教学实践就能成为一名好教师。

在教学改革中我也曾提出研究性教学的六个要素，即：① 知识体系具有完备性；② 教学进程符合认知规律；③ 教学内容极具启发性；④ 实验与理论教学配合紧密，具有研究探索因素；⑤ 教学环节张弛有度、赋予学生较大的自我发挥空间；⑥ 能够最大限度地调动学生积极性、想象力、创造性。同时注意教学中的六个环节，即：① 理论教学；② 实验教学；③ 科学的讲授艺术；④ 现代化教学手段的运用；⑤ 科学研究进展的引领作用；⑥ 教材建设。把握好这六个要素及教学的六个环节就一定能做好教学工作，实践证明教学效果也很好。这些经验我希望对青年教师有一定的启发作用。

问：现在高校和企业都在人工智能上展开了很多的研究，而且企业的计算资源可能比高校更丰富。特别是在深度学习出现之后，对于数据和计算资源的要求越来越高，造成高校在做研究的过程中会遇到一些计算瓶颈，特别是一些小规模的研究团队。您觉得对于高校研究人员而言，如何应对这一状况呢？

阮教授：科学研究是高等院校的另一个主要任务，但高

等院校的科研任务与企业还是有差别的。我历来主张“企业做今天的事，高校做明天和后天的事，企业要与高校密切合作，各取所长，这样才能高效快速地促进科学技术的发展，并能使高校的科研成果快速落地，尽快形成生产力。”在国外，尽管企业有很强的研发能力，如贝尔实验室这样的企业实验机构，它们也十分重视与高校的合作。现在高校与企业也建立了合作基地，联合培养人才，合作开发项目，但还不够深入，还有进一步提升的空间。在基础理论研究方面国家基金委、科技部、教育部及省市都有相关支持，但企业中也有大量的难题需要高校参与才能解决。科研立项要遵循两大原则，即：1. 需要性原则——社会或经济发展的需要；2. 科学性原则——要有科学的理论、事实为依据；前者企业提出，后者高校来求证和给出答案。这两者是密不可分的。一般来说高校的理论优势是不太容易被取代的，关键是高校科研人员要放下身段，与企业紧密合作；高校政策上要把“纵向”与“横向”项目一视同仁。在这方面我们确实存在短板，一方面国家积极支持，企业要有长远规划，积极投入，在研发上要注重企业发展的后劲；另一方面高校要深入现场，了解需求，特别是以应用科学为研究目标的工科院校更应如此，建立联合实验室可能是一个有效的途径。

问：您还在多个学会或者学术组织中担任重要职务，这些经历对您开展教学科研工作有哪些帮助？在组织相关科研活动或者学术会议过程中有哪些趣事可以跟我们分享么？

阮教授：我在学术团体中任职较多，如电子学会常务理事，学术委员会委员、会士；中国通信学会理事、会士；中国图像图形学会理事，奖励委员会副主任；IEEE 北京分会主席；IET 北京分会主席、Fellow；通信学会信号处理学会副主任；信号处理学会副主任委员、中国电子教育学会理事，研究生分会常务理事等等。参加学会工作主要是结识同行，学术交流，为学校学科建设做贡献，同时提高我们学科的知名度。而且个人在学术上也获益匪浅。现在为了年轻教师尽快成长，多个学术团体中的

图 9 阮秋琦教授（左）主持国际会议

任职已转移给年轻教师了，保证了学校的学术地位和声誉上的传承，有利于学校的双一流建设。

问：您怎么看目前计算机视觉领域的发展现状？有哪些优势和不足？

阮教授：计算机视觉研究的目标有两个：1）、自动构造场景描述的图像理解系统；2）、理解人类视觉，以便有朝一日用机器代替人去做人类难以达到或根本无法达到的工作。

20 世纪 80 年代，在计算机视觉研究中占主导地位的是 Marr 教授提出的视觉计算理论框架。在这种框架下，Marr 认为计算机视觉可看作是三个层次的信息处理过程，而且要从计算理论、算法描述及硬件实现三个方面去实现三个层次的工作。也就是他提出的 a) 初始简图（primal sketch）；b) 2.5 维简图；c) 三维模型表示。20 世纪 90 年代，Rosenfeld 认为在计算机视觉研究中应重视三个方面的工作：1） 计算的鲁棒性问题；2）主动视觉（active vision）的研究；3）定性视觉的研究（qualitative vision）。

综合目前计算机视觉领域的研究成果来看，在理论上，计算机视觉的研究尚未形成自己的理论体系。在技术上，在许多方面还达不到实际应用的要求；在功能上，与人的视觉相比仍处在低水平阶段。计算机视觉研究突破性进展在很大程度上依赖于人对自身视觉机理的深

入了解。尽管如此，国内外的研究工作已取得了大量的成果。在机器人视觉及车辆自动导航方面的成果是令人乐观的。在三维描述精度上显然高于人的视觉。

计算机视觉研究在过去的 50 多年中取得了长足的进展，其研究内容和应用前景都早已超出了研究者的初衷。尤其是计算机技术的迅速发展对计算机视觉的研究起到了推波助澜的作用。人赋予机器以视觉能力，进而发展智能机器人出现了新的研究热潮，各国都相继投入了大量的人力、物力、财力，希望在计算机视觉研究中能够取得突破。今后的工作我认为应集中在如下几点：

（1） 提高处理的鲁棒性。

（2） 在信息获取方面开展弱定标或自定标研究，以利于实用性系统的开发。

（3） 加强边缘学科的研究工作，如：人的视觉特性、心理学特性等的研究如果有所突破，将对计算机视觉技术的发展有极大的促进作用。

（4） 重视开发那些专门的和部分模拟视觉的系统。这主要集中在三维物体与自然景物的识别与分析上，不但是孤立的图像识别，而且应具有一些推理及联想能力。这样，有可能进一步推动计算机视觉研究的发展。

问：对从事计算机视觉领域研究的青年学者和学生有没有寄语？

阮教授：希望广大青年学者“青出于蓝而胜于蓝”。

责任编委 明悦 张军平 贾熹滨

阮秋琦

阮秋琦教授 1969 年毕业于北方交通大学并留校任教。1978 年 8 月至 1981 年 12 月为北方交通大学通信与控制工程系首届研究生；1981 年 12 月于北方交通大学研究生毕业获中国首届工学硕士学位；1985 年 1 月至 1986 年 1 月国家公派赴美国匹兹堡大学访问进修，主修图像处理与计算机视觉；1987 年 2 月至 1990 年 9 月国家公派美国及辛辛纳提大学进修并做客座研究工作，主攻图像处理和计算机视觉；1991 年回国任北方交通大学教授；1992 年至今任北方交通大学二级教授及博士生导师。1994 年赴美国伯兰戴斯大学做客座教授；1996 年 3 月至 1997 年 9 月再次赴美国匹兹堡大学任客座教授。阮秋琦教授曾任信息研究所所长、通信与控制工程系主任、电子信息工程学院院长、计算机与信息技术学院院长、软件学院院长、国务院学位委员会学科评议组成员。现任校务委员会委员，IET Fellow、IET 北京分会主席，IEEE 终身高级会员、北京分会主席，中国通信学会会士、中国通信学会信号处理分会副主任委员，中国电子学会会士、理事、技术委员会副主任、中国电子教育学会理事、研究生教育分会常务理事，“北京技术预见行动计划(信息技术领域)”专家组成员，信号处理学会副主任委员，国家自然科学基金委员和科技部国际合作项目评审专家，国家留学基金委员会评审专家、清华大学等兼职教授。阮秋琦教授曾承担国家自然科学基金重大和面上项目、国家 ″863″、“973”、铁道部以及省、市级科研项目 80 余项。发表论文 490 余篇，出版书籍 6 部，译著 7 部，主编“中国铁路百科全书通信信号卷“等 3 部，获国家专利 3 项。阮秋琦教授作为信号与信息处理学科学术带头人，其撰写的论文多次获得铁道部、专业学会优秀论文奖。其著作曾获校优秀教材特等奖、铁道部优秀教材二等奖，曾获北京市科学技术二等奖、国家教委科技进步二等奖、铁道部科技进步一、二、三等奖，北京市优秀教学成果二等奖（2 项），北京市教学成果一等奖（3 项），国家级教学成果二等奖以及专业学会精品教材、优秀教材、国家级精品课、国家级资源共享课，北京市翱翔计划优秀指导教师等多项奖项。被评为国家级有突出贡献的专家、国家级教学名师、中组部联系的党内高级专家、北京市教学名师、铁道部有突出贡献专家、铁道部优秀科技工作者、北京市优秀教师、詹天佑北方交通大学奖，茅以升科技奖、詹天佑科技人才奖、宝钢优秀教师特等奖，并享受国家政府津贴。

纪念孙剑老师

2022 年 6 月 14 日临晨，计算机视觉领域传来噩耗：旷视研究院院长孙剑老师因突发疾病抢救无效离世，享年 45 岁。

孙剑老师是计算机视觉领域具有重要国际影响力的一位学者，1976 年出生于西安，分别于 1997 年、2000 年、2003 年在西安交通大学获得了学士、硕士、博士学位。2003 年博士毕业之后加入微软亚洲研究院，从事计算机视觉与计算机图形学的研究；于 2016 年加入旷视科技，担任首席科学家与旷视研究院负责人。2019 年，担任西安交通大学人工智能学院首任院长。孙剑老师的主要研究兴趣包括：计算摄影学与基于深度学习的图像理解[1]。

一、华人学者先锋，两获 CVPR 最佳论文奖

孙剑老师在计算机视觉领域取得了多项具有重要国际影响力的开创性学术成果。在他一生的研究贡献中，重要的贡献包括但不限于：

- 带领团队提出著名的深度残差网络 **ResNet**，成功地解决了深度神经网络训练难的世界级难题，使得超过 100 层的深度神经网络的训练成为现实，是计算机视觉领域的重大突破与深度学习技术的重要里程碑[2]。成功应用于多项计算机视觉任务，成为为数不多的华人学者提出的计算机视觉“标配”主干网络之一。该项工作获得了 2016 年 CVPR 会议最佳论文奖。
- 带领团队发现图像的暗通道先验（Dark Channel Prior），巧妙地将图像去雾问题转化为 matting 问题，并成功应用于单幅图像去雾。该项工作获得亚洲第一个 CVPR 最佳论文奖（2009 年）。

CVPR 2022 大会现场，在公布最佳论文之前，组委会专门播放了一段视频缅怀孙剑老师。除了上述两个 CVPR 最佳论文奖，孙剑老师还获得过多项重要科研奖励[3]，包括但不限于：

- 2020 年获得 AI 2000 计算机视觉全球最具影响力学者第二名
- 2019 年获得何梁何利基金“青年创新奖”
- 2017—2019 年带领团队获得 MS COCO 物体检测世界比赛三连冠
- 2017 年带领旷视研究院击败谷歌、Facebook、微软等企业，获得 COCO&Places 图像理解国际大赛三项冠军（COCO 物体检测、人体关键点，Places 物体分割）
- 2016 年获得国家自然科学二等奖（视觉场景理解的模式表征与计算理论及方法）
- 2015 年获得高等学校科学研究优秀成果自然科学一等奖视觉场景理解的模式表征与计算理论及方法）
- 2015 年，带领团队获图像识别国际大赛五项冠军（ImageNet 分类，检测和定位，MS COCO 检测和分割）
- 2010 年获评《MIT Technology Review》35 岁以下科技创新 35 人

二、投身 AI 产业化，掌舵旷视研究院

2016 年，孙剑老师加入旷视科技，担任首席科学家与旷视研究院负责人。在孙剑老师带领下，旷视研究院研发了包括移动端高效卷积神经网络 ShuffleNet、开源深度学习框架天元 MegEngine、AI 生产力平台 Brain++等多项创新技术[3]。

拒周少华老师回忆：在 CVPR 2017 视觉之友会议上，孙剑老师针对其加入旷视科技之后的经历做出了一个重要发言[4]：”做学术研究首先要讲创新，但做应用研究或产品可能有些人觉得做到能用就行了。我最近一年的深刻体验是一款产品或一种服务，必须要有某种创新型，才有可能在市场竞争中获得先发优势或竞争力。没有创新，没有差异化，简单 copy 别家 idea 的产品可能都走不出家门。举个例子，人脸识别技术是有很大的突破，但是只有结合新的场景才能在市场做出好的产品。Face++的“Paying with your face”在中国第一个把基于人脸识别的身份认证应用于金融行业，打造了一款成功的在线云服务，今年也被 MIT TR 评为 2017 年全球十大突破性技术。由创新带来的先发优势和更深入的产品迭代，并不是后来者可以轻易 copy 成功的。”

三、教书育人，帮助青年人才

2019 年，孙剑老师担任西安交通大学人工智能学院首任院长在人工智能领域为西安交通大学培养了人才，曾培养出深度学习博士张祥雨等人工智能领域研究人员[3]。

孙剑老师对博士毕业的选择：1) 首先博士毕业后的 2-3 年是最关键的，这段时间要做的更强、为今后打好基础；2) 做好找一个周围同事水平高于你的环境，持续 push 自己变得更强；3）公司研究院往往高手的密度更高、合作比较容易，同时能够 100%专注理解和研究难的问题[4]。

据清华大学鲁继文老师回忆：2017 年 11 月份，当时西安交通大学郑南宁院士团队组织申报一个 2017 年新一代变革性技术国家重点研发计划，孙剑老师作为项目负责人，鲁继文作为课题负责人，大家一起准备申报书。申请的项目是下一代深度学习理论、方法与关键技术。虽然当时大家都在用深度学习来解决计算机视觉中的各种问题，但是对深度学习理论方法与关键技术的理解其实不是很深刻。孙剑老师与课题组进行了多次讨论，最后创新性地提出了“动力学原理驱动的深度学习基础理论与技术”，当时课题组都感觉这个思路眼前一亮，将深度网络看作一个动力学系统进行优化，非常具有启发性。2018 年 4 月份项目进行答辩汇报，课题组在西安交大集中修改 PPT。孙剑老师一直在企业界，对学术界的答辩 PPT 不是十分擅长，然后他多次把自己封闭在宾馆房间进行彩排演练，最后项目也顺利申报成功。

2018 年参加 ECCV 国际会议，在从慕尼黑回北京候机的时候，孙剑老师听说鲁继文在开始研究机器人抓取，尤其是利用视觉感知进行机器人抓取的操作，很感兴趣。说可以支持这项研究，最后从旷视科技资助了一个金额不小的项目，希望能把机器人抓取这个方向做好。现在鲁继文老师实验室的两台机械臂，就是用孙剑老师当时资助的项目购买的。

鲁继文老师这几年因为国家重点研发计划项目合作的机会，多次跟孙剑老师一起开会，讨论合作项目的中期检查、年度检查等各种材料，每次都能从孙剑老师那里学到很多。因为清华大学和旷视科技都在北京，过去这几年在国内外出差途中鲁继文老师多次跟孙剑老师遇到同一个航班或者同一趟高铁，每次都能看到孙剑老师在旅途中拿出一本书静静地阅读。

据前旷视科技实习生章圳黎回忆：孙剑老师给他的最大感受是水平非常高，但仍然非常谦和，平易近人，其中最触动章圳黎的一点是虽然已身居高位而且日常工作繁忙，孙剑老师依然对新的工作和论文跟得非常紧每天早上他基本是研究院最先来公司的一批人，然后一

般会花几十分钟看一下最新的论文，碰到有趣的工作就会转给整个研究院。还有一次开 Valse 听报告，章圳黎碰巧坐孙剑老师后面，开场行政部分没啥有用的信息，章圳黎看孙剑老师在用手机看论文，孙剑老师水平那么高了尚且这么拼，作为后辈真的没有理由不更加努力。

孙剑老师对自己要求很严，但对研究院的同学却非常体谅和宽容，这个不是说对同学们要求很低，而是说如果感觉到同学们真的努力了，即使最后结果没有特别好，孙剑老师通常也是鼓励大于批判。章圳黎在旷视科技实习期间研究的论文，当时截止日期之前，孙剑老师仍然抽空帮他改了论文，而最后论文中了之后，孙剑老师碰到章圳黎说“不错，第一次投就中了，继续努力”。

愿孙剑老师一路走好！

参考文献

[1]. 孙剑老师个人主页：http://www.jiansun.org/

[2]. 学院、研究所简介-西安交通大学-人工智能学院（xjtu.edu.cn)

[3]. 百度百科词条：孙剑（旷视原首席科学家、旷视研究院原院长）

[4]. 周少华：Rest In Peace！ResNet In Use！| 学术人生（视觉求索公众号）

[5]. 鲁继文：纪念孙剑老师

[6]. 章圳黎：纪念孙剑老师

责任编委 **金鑫**

COMPUTER VISION NEWSLETTER

03 2022
总第 33 期

计算机视觉专委会简报

CCF 计算机视觉专委会