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 2021

总第 29 期

CCF 计算机视觉专委会

COMPUTER VISION NEWSLETTER

计算机视觉专委会简报

2021 年第 03 期

总第 29 期

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 103 期 贵州师范大学

2021 年 7 月 8 日上午 9 时由中国计算机学会计算机视觉专委会主办、贵州师范大学大数据与计算机科学学院承办的第 103 期 CCF-CV 走进高校系列报告会在贵州师范大学花溪校区学术会议中心成功举行。本次会议邀请了北京大学彭宇新教授、杭州电子科技大学俞俊教授、中科院自动化所赫然研究员、西安电子科技大学王楠楠教授做特邀报告。会议由贵州师范大学大数据与计算机科学学院院长胡圣波教授、欧卫华教授共同担任执行主席。专家们围绕“计算机视觉前沿技术及应用”做了精彩报告，引起了师生及同行的广泛共鸣。

此次 CCF-CV 走进高校系列报告会专家们的讲解深入浅出，从实际出发分享自己学术上的经验，研究点评环节专家们的指导切中要害，为切实提升青年教师的研究能力提供了帮助！希望通过 CCF-CV 走进高校系列报告会让计算机视觉学科的专业内容更好地开展交流、发展战略研究，促进国内学者间的了解与合作，推动国内计算机视觉学科发展，提升我国计算机视觉研究在国际领域的影响力。

第 104 期 湖北民族大学

2021 年 7 月 30 日下午，由中国计算机学会计算机视觉专委会（CCF-CV）主办、湖北民族大学承办的第 104 期 CCF-CV 走进高校系列报告会，以线下参会+线上同步直播的形式在湖北民族大学成功举办。本次活动邀请了北京大学林宙辰教授、重庆邮电大学李伟生教授、中国科学院自动化研究所张兆翔研究员、中国科学院信息工程研究所任文琦副研究员四位专家学者做特邀报告。湖北民族大学信息工程学院院长谢坤武教授担任报告会的执行主席和主持人。

此次 CCF-CV 走进高校系列报告会聚焦领域前沿，四位专家以通俗易懂和深入浅出的方式为参会人员带来精彩的报告分享。在研究点评环节专家们对两位青年教师的研究工作给予了肯定，并提出了中肯建议，为他们后续研究工作开拓了思路。最后，湖北民族大学谢坤武教授再次向各位特邀专家和现场参会人员表示感谢，并欢迎计算机视觉领域的同行专家学者来湖北民族大学大学指导和交流工作，报告会在全场热烈的掌声中圆满结束！

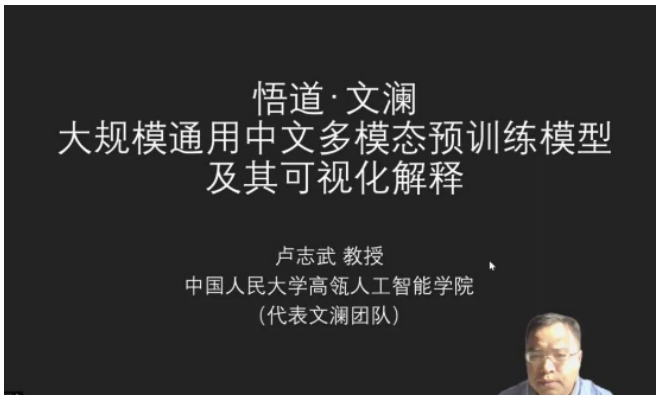
责任编委 毋立芳

第 10 期 视觉与语言（Vision & Language）的前沿进展与未来趋势

CCF-CV 视界无限系列研讨会

2021 年 8 月 29 日，由中国计算机学会计算机视觉专委会主办的第 10 期 CCF-CV “视界无限”系列活动——“视觉与语言（Vision & Language）的前沿进展与未来趋势”研讨会通过线上方式成功举办。研讨会邀请了中国人民大学卢志武教授、杨征元博士、中科院自动化所黄岩副研究员、西北工业大学王鹏教授和美团智慧交通平台视觉智能部马林研究员做主题报告并参与圆桌讨论。中国计算机学会计算机视觉专委会主任、北京大学查红彬教授出席活动。本期研讨会由北京航空航天大学人工智能研究院承办，刘偲副教授任执行主席并主持会议。

中国计算机学会计算机视觉专委会主任、北京大学查红彬教授首先致辞。查教授对参会的各位老师及同学表示欢迎。他指出视界无限系列研讨会是中国计算机学会计算机视觉专委会举办的品牌活动，旨在促进同行之间的相互交流并针对具体的计算机视觉问题进行深入探讨。查教授表示，本次研讨会的主题“视觉与语言”涉及到了智能信息处理当中两个不同模态方面的研究，视觉处理通常是针对一些低层处理，而语言更多是高层次的处理，如果能够将二者很好地结合，使它们在不同层面进行融合或交互，就能够帮助我们找到人工智能处理当中的一些有效办法，为研究提供新的思路。希望本次研讨会能对从事这一领域研究的老师和同学们有所启发。最后代表专委会感谢主办单位各位老师同学为筹办本次研讨会作出的努力，预祝研讨会圆满成功！



中国人民大学卢志武教授的报告题目是“大规模通用中文多模态预训练模型及其可视化解释”。首先，卢老师介绍了中文多模态预训练面临的数据收集难题以及解决办法，即从互联网上爬取海量图文数据。其次，卢老师介绍了在爬取的弱相关图文数据上设计多模态预训练模型-文澜，并提出了基于 DeepSpeed 的预训练

算法。最后，卢老师通过下游任务评测以及神经元可视化展示了文澜强大的理解能力。

杨征元博士报告的主题是“Visual grounding: Building Cross-Modal Visual-Text Alignment”。他从两个层面定义和讨论了 visual grounding，分别是狭义的 visual grounding 任务以及广义的跨模态表征学习。在第一部分，针对狭义的 visual grounding 任务，他介绍了一种 one-stage grounding 方法，大大增加了模型的运算速度和计算精度。他同时也讨论了如何将这一方法进一步优化，延伸至弱监督设定，以及视频任务。在第二部分，他介绍了如何将 visual grounding 与视觉语言任务结合。分享了使用预训练以及目标对应损失函数等学习 visual grounding 的方法，及其在视觉问答，captioning，点云定位等任务上的应用。

中科院自动化所黄岩副研究员的报告题目是”图文匹配研究进展”。首先，他回顾了图文匹配的发展历程，并指出跨模态语义鸿沟和跨模态少样本等是该任务目前所面临的主要挑战。然后，他针对跨模态少样本问题，提出了跨模态长时记忆网络，能够选择性存储和更新成对跨模态小样本特征，以知识复用的方式强化少样本图像和文本之间的关联性。最后，他简要展望了未来研究趋势，并指出细粒度跨模态对齐和图文匹配模型小型化是比较有潜力的研究方向。

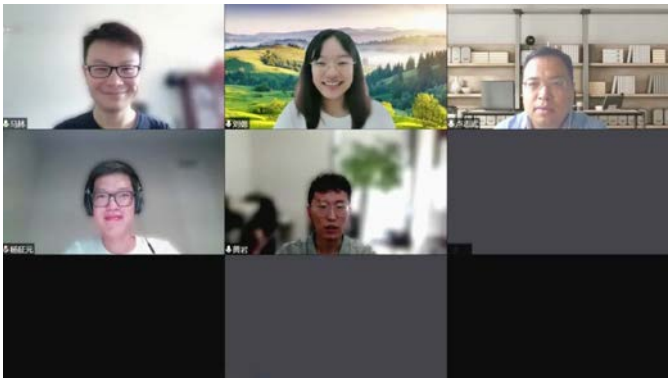
西北工业大学王鹏教授的报告题目是“Richer and Deeper: Vision and Language Understanding with Richer Visual Content and Deeper Non-visual Knowledge”。主要介绍了他在视觉—语言领域两个方面的工作。首先是如何充分挖掘视觉信息中丰富的语义信息，特别是文本信息，从而帮助提升模型的视觉理解与推理。随后他介绍了如何将视觉信息以外的人类知识引入视觉语言模型，从而帮助机器更聪明的理解人类的语言指令。

美团智慧交通平台视觉智能部研究员马林的报告题目是“Vision+Language: From Captioning to Grounding（视觉与语言结合的研究）”。首先马林介绍了视觉与语言相结合研究上的一些挑战。主要挑战在于文本是结构化的信息，而图像和视频主要是非结构化的信息。如何将结构化的信息和非结构化的信息融合起

来共同学习视觉和文本的联合信息是比较大的挑战。随后马林介绍了其团队在描述生成，视频定位，语言指代图像分割等方向的一些研究成果。最后，通过介绍相应的研究工作，马林认为如何设计深度模型来挖掘图像/视频与文本之间的交互关系对解决视觉和文本相结合的任务尤其重要。

Panel 实录

紧接着是 panel 环节，由北京航空航天大学刘偲老师主持，与各位讲者探讨了视觉与语言的前沿进展与未来趋势。整个研讨会在中午 12 点圆满结束。



刘偲：大家好，我们今天主要围绕着视觉语言这个主题与各位老师展开探讨。首先请各位老师聊一聊在 visual language 领域，学术界和工业界的差距主要在哪？未来在工业界的落地点和应用场景有哪些？如何打破学术界和工业界的 gap，将 visual language 推进到实际落地？

马林：首先在真正落地时，面临的现实场景比已有的 benchmark 数据集合复杂很多。所以在工业上应用时，数据集会限制整个能力的提升。第二点，工业界还是有很多应用场景，但因为 visual language 偏底层，很多时候用户并非直接体验到这个技术。能看到的几个点：比如图像及短视频这种信息流的产品，或是对图文信息进行推荐跟搜索时，也会用到大规模预训练模型来操作，只不过它在底层。另外大家以图搜图的时候，比如拍照购物的场景，商品不仅有图像信息也有文本信息，二者融合也能帮助以图搜图的操作。还有对话系统，如果要想变成非常自然的语言场景，也是多模态的，因为大家在聊天过程中也会发图片。还有一些未来应用场景。比如监控，想搜到某个时间点穿什么衣服的人，用自然语言来搜是最常用的。但是目前在业界，技术跟数据集可能还不够充分，所以这方面还有些 gap，但整个 gap 是在不断缩小的。

卢志武：首先在文澜 2.0 出来以后，华为特别希望在图文检索上面把这种多模态预训练的东西用进去，我们也的确看到了很好的效果。第二是文澜 3.0 想做的事情：视频加文本的预训练，把它做好对视频推荐功能也会有很大帮助。还有就是各种用户创作，也是很有潜力的一点。

黄岩：刚才提到的跨模态创作，比如用语言生成图像视频，目前我所了解到的有不少公司对此技术有较大需求。另一块也和创作相关，像文章自动配图和新闻自动配图。至于 gap，应该是说如何在大规模数据有噪声的情况下让模型优化得更好。

刘偲：好的，谢谢各位老师。其实近几年视觉语言模型在落地中确实有了很大进展，但相比于单一模态，还会存在哪些优势或者瓶颈？以及应该如何在多模态预训练中合理使用带噪声的数据呢？

马林：我觉得相比单一模态整体还是更具优势的。首先数据往往还是多模态的情况：无论在互联网或其他场景，基本是短视频信息流或图文以及各种评论信息。如果模型设计的好，会把信息吸收进来帮助整个模型提升，或者如果它真的是噪音，也可以过滤掉。所以怎样更好地利用噪音数据是一个重要的点。

杨征元：总体是有优势的，如果真说劣势，可能一些传统模型在 input 模态的地方不够灵活，但这更多是模型的问题，不是多模态本身的问题。

黄岩：我接触的实际需求更多是做多模态融合，比如以前做单纯基于视觉的效果已经挺好，但把文本、语音等其他模态用上，可以做得更好。另一个情况是，比如一种模态在某些情况下表现不好，而另一个模态在该情况下很稳定，因此还可以利用多模态之间的互补性。另外，关于噪声的数据，大家可能更多地倾向于使用主动学习或者是数据挖掘的方式，机器学习领域有很多研究人员在探索这一块。但是如何放到多模态场景下做得比较好，目前好像还没人去考虑这部分。

卢志武：我们做多模态的时候，面临第一个问题就是爬数据，因为现存的中文数据太少了。我们观察到，中国人说话很含蓄，可能转了两三道弯才能把图文联系上。所以在对数据做完一些简单的清洗后，觉得这些东西应该留着，最后在 6.5 亿图文对里面只筛掉很少敏感数据。在我们的多模态模型训练完以后，再去分析它就会发现这是个好事情，比如涉及到抽象理解，我们的模型有优势了，而这个能力其实就是从数据中学来的。所以有时候我们大家认为的噪声不一定是噪声。

刘偲：我们进入下一个问题。利用 unpaired 视觉语言数据进行无监督训练得到的模型会不会有 bias？如果存在，目前有哪些方式可以缓解 bias 的问题？

卢志武：肯定会面临这个问题，但我们的处理方式是很简单的。相当于永远先把 unpaired 数据当成单模态的，每个单模态先训练好，最后拿 paired 的数据去把跨模态那部分数据训练好，大概做法就是这样的。甚至在跨模态里面也可以把那些单模态自监督的 loss 加上，目前来看效果还是挺好的。

刘偲：下一个问题是多模态预训练模型是不是可以提升对单模态的理解？如何通过多模态预训练模型提升单模态特征的表示能力？

卢志武：这要通过实验证明，比如说在文本模态上，同为文本 encoder，通过单模态和多模态预训练得到的模型进行对比，如果多模态训练效果好才能证明它是有价值的。所以我们做了中文新闻分类实验，初步证明了多模态的效果。同时我们也和自动化所一位老师合作，想通过脑机接口进一步回答这个问题，目前初步得到一些结论：通过多模态预训练的文澜，在类脑指数上确实要比单模态训练高。

马林：我还是从视觉跟语言两个模态不在一个 level 上来说。从 NLP 角度来看，文本其实相当于比较高级抽象的语义信息。中文或英文的词表其实是比较小的，比如中文常用词可能大约是万级别的，到字来说就几千，每个字和词的语义不一样而且是高度抽象的。而视觉信息就偏底层，每个 pixel 其实没有语义信息。所以我觉得多模态预训练的模型，对视觉上面的提升更多一些。假如纯 NLP 的数据，它的广度以及词汇数据量对于 tonken 来说已经非常多了，但是对于视觉来说，整个视频或者图像，结构也还是比较少的。

杨征元：我看到多模态首先第一点：同是 visual tasks，对于那些相对比较 long-tail 的部分或者训练的比较少想要一般化的时候，他会对性能和收敛的速度有好处。比方 language 对 visual，language 有更多标签以外的东西，所以多少会看到一些 long-tail 的东西，也会有帮助。第二点，我之前看到更多的帮助也是在 visual 那端。但如果有一些特殊任务，或者在一些情况下对 NLP 有帮助的话，那这个事情其实挺有启发性的。

刘偲：下一个问题是目前这个视觉语言模型大家都认为是从感知往认知方向的过渡，它有一定的认知但和跟人类相比还是有一些差距。各位老师认为，如果要达到人类认知的水平，差距主要在哪里，未来该如何弥补？

杨征元：我觉得这个问题前半部分比较好答，大家都知道有这样的问题，但至于未来怎么做，还是有挑战的。比如对于大模型，给它 few shots 的例子是否真的能 adapt。这一点 GTP-3 展示了一些，但还是存在差距。另外，对于 GTP 或者文澜这样的大模型，它可能是去年训练好的，不知道今年的信息，要怎么样合理地把它加进去也是一个重要的问题。最后还是回到 data。模型感知的 data 质量和人不是完全一致。所以总结起来，第一步还是需要更好地理解现有这些很强的模型，然后下一步才是基于我们的理解做一些事情。

马林：现在学的更多是数据共生的 pattern，其实跟人类认知的差距还是挺大的。人类对 few shots 和 one shot 的理解能力还是蛮强的，预训练模型现在还不能完成这样的任务，尤其是需要强推理时，或是对脑的认知方面。第二点，至于未来如何弥补。现在很多做 symbolic reasoning，这也是对 knowledge 的一个表示方式，可能是一个未来。这好比从人的角度来说是举一反三，从模型的角度来说是有历史规律的总结。再者如果能对可解释性做更好的挖掘，也是一种方法。

黄岩：我当时跟做生物和心理的研究人员交流过，

他们推荐我一本书叫《认知心理学》。虽然是心理学领域的书，但主要是在介绍什么是认知机制或者认知过程，并由浅入深地包括视觉注意、记忆、长短时记忆，还有相应的知识推理、决策等。这些认知机制或者过程跟现在一些模型是很相似的。像 transformer 或 attention，就是对视觉注意进行建模；像 memory、神经图灵机，或者长短时记忆网络也是受到记忆机制建模启发；像 symbolic reasoning 就是关于推理机制建模；现在我们做的强化学习，其实更多是跟决策关联在一起的，它们本身就是人类的认知机制从底层到高层的演变过程。所以一个可能思路是，在基础网络框架之上去建模多种认知机制，包括视觉注意，记忆推理，决策等，即深度认知学习。

刘偲：最后一个问题是当前基于大规模数据预训练的视觉语言模型取得了优异的性能，而类似 symbolic reasoning 一类的推理方法的性能却不尽如人意，那未来基于推理的方法会不会逐步被预训练所取代？我看到有一些方法，其实也已经把知识图谱结合在预训练模型当中了，结合知识的预训练模型，是不是可以呢？

杨征元：现在一些文章说大规模预训练模型在之前 reasoning benchmark 上已经做的比 reasoning 的方法好了，但我还一直期待能看到究竟要怎么样去理解这些东西，有时候可能只是通过数据找到了一条 short cut，刚好适配了 reasoning 的 testing set。

卢志武：目前一些观点觉得预训练就是 over fitting 或是模板学习，我其实不太认可这个观点。我们做了很多的可视化分析，比如我做报告使用的遥感例子，对于棒球场俯视的概念，我们预训练数据里是没有的，而我们模型是真正学到这个概念之后推广到棒球场上去了，这也算是认知能力的一种，甚至可以说是某种推理能力。

责任编委 杨巨峰

CCF-CV 秘书处 2021 年度第二次工作会议顺利召开

2021 年 7 月 14 日，中国计算机学会计算机视觉专委会（CCF-CV）秘书处本年度第二次工作会议在天津召开，专委副主任王亮研究员出席会议并指导工作，秘书长马占宇教授主持会议。本次会议主要讨论如何落实专委会常务委员会第七次会议的各项决议，并制定秘书处下一阶段的工作计划。

上半年，专委各项学术活动和组织建设稳步推进，走进高校、视界无限、视觉前沿讲习班等特色活动成功举办并持续扩大影响力，展示了 CCF-CV 专委旺盛的组织活力。接下来，如何办好专委年度学术论坛 RACV 2021 和年度全体委员工作会议、如何更有效地激发海外华人学者参与专委活动的积极性，大家围绕这些议题展开了热烈讨论，并形成了具体可行的执行方案。

责任编委 黄岩

专题综述

计算机视觉中拥抱 Transformer 的五个理由

微软亚洲研究院 胡瀚

“统一性”是很多学科共同追求的目标，例如在物理学领域，科学家们追求的大统一，就是希望用单独一种理论来解释力与力之间的相互作用。人工智能领域自然也存在着关于“统一性”的目标。在深度学习的浪潮中，人工智能领域已经朝着统一性的目标前进了一大步。例如对于一个新的任务，基本都会遵循同样的流程对新数据进行预测：收集数据，做标注，定义网络结构，训练网络参数。但是，在人工智能的不同子领域中，基本建模的方式各种各样，并不统一，例如：自然语言处理目前的主导建模网络是Transformer；计算机视觉很长一段时间的主导网络是卷积神经网络（CNN）；社交网络目前的主导网络则是图网络等。

尽管如此，从去年年底开始，Transformer 还是在 CV 领域中展现了革命性的性能提升。这就表明 CV 和 NLP 有望统一在 Transformer 结构之下。这一趋势对于两个领域的发展来说有很多好处：（1）使视觉和语言的联合建模更容易；（2）两个领域的建模和学习经验能深度共享，从而加快各自领域的进展。

一、Transformer在视觉任务中的优异性能

视觉 Transformer 的先驱工作是谷歌在 ICLR2021 上发表的 ViT[1]，该工作把图像分成多个图像块（例如 16x16 像素大小），并把这些图像块比作 NLP 中的 token。然后，直接将 NLP 中的标准 Transformer 编码器应用于这些“token”，并据此进行图像分类。该工作结合了海量的预训练数据，例如谷歌内部 3 亿图片分类训练库 JFT-300M，在 ImageNet-1K 的 validation 评测集上取得了 88.55%的 top-1 准确率，刷新了该榜单上的记录。

ViT 应用 Transformer 比较简单直接，因为其没有仔细考虑视觉信号本身的特点。所以，它主要适应于图像分类任务，对于区域级别和像素级别的任务并不是很友好，例如物体检测和语义分割等。为此，学术界展开了大量的改进工作。其中，Swin Transformer 骨干网络[2]在物体检测和语义分割任务中大幅刷新了此前的记录，让学界更加确信 Transformer 结构将会成为视觉建模的新主流。具体而言，在物体检测的重要评测集 COCO 上，Swin Transformer 取得了单模型 58.7 的 box mAP 和 51.1 的 mask mAP，分别比此前最好的没有扩充数据的单模型方法高出了+2.7 个点和+2.6 个点。此后，通过改进检测框架以及更好地利用数据，网络的性能进一步取得了 61.3 的 box mAP 和 53.0 的 mask mAP，累计提升达+5.3 box mAP 和+5.5 mask mAP。在语义分割的重要评测数据集 ADE20K 上，Swin Transformer 也取得了显著的性能提升，达到 53.5 mIoU，比此前最好的方法高出 3.2 mIoU，此后随着分割框架和训练方法的进一步改进，目前已达到 57.0 mIoU 的性能。

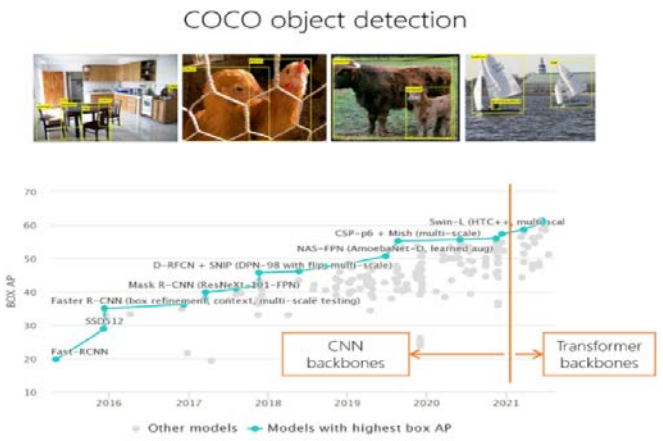


图 1　历年 COCO 物体检测评测集上的纪录

除了在物体检测和语义分割任务上表现优异外，基于 Swin Transformer 骨干网络的方法在众多视觉任务中也取得了优异的成绩，例如视频动作识别[3]、视觉自监督学习[4]、图像复原[5]、行人 Re-ID、医疗图像分割[6]等。

Swin Transformer 的主要思想也比较简单直接，就是将具有很强建模能力的 Transformer 结构和重要的视觉信号先验结合起来。这些先验主要有层次性（Hierarchy），局部性（locality）以及平移不变性的特点（translation invariance）。Swin Transformer 的一个重要设计是移位的不重叠窗口（shifted windows）。不同于传统的滑动窗，不重叠窗口的设计对硬件实现更友好，从而具有更快的实际运行速度。如下图左所示，在滑动窗口设计中，不同的点采用了不同的邻域窗口来计算相互关系，这种计算对硬件不太友好。如下图右所示，Swin Transformer 使用的不重叠窗口中，统一窗口内的点将采用相同的邻域来进行计算，对速度更友好。实际测试表明，非重叠窗口方法的速度比滑动窗口方法快 2 倍左右。在两个连续的层中，还做了移位的操作。在 L 层中，窗口分区从图像的左上角开始；在 L+1 层中，窗口划分则往右下移动了半个窗口。这样的设计保证了不重叠的窗口间可以有信息的交换。

图 2　传统的滑动窗口方法（左），由于不同的查询所用到的关键字集合不同，其对存储的访问不太友好，实际运行速度较慢。移位的不重叠窗口方法（右），由于不同的查询共享关键字集合，实际运行速度更快，从而更实用

在过去大半年中，学术界视觉 Transformer 还涌现了大量变种，包括 DeiT[7]，LocalViT[8]，Twins[9]，PvT[10]，T2T-ViT[11]，ViL[12]，CvT[13]，CSwin[14]，Focal Transformer[15]，Shuffle Transformer[16]等。

二、拥抱Transformer的五个理由

除了刷新很多视觉任务的性能记录以外，视觉 Transformer 还拥有诸多好处。事实上，过去 4 年学术界不断挖掘出 Transformer 建模的各种优点，可以总结为如下图所示的五个方面。

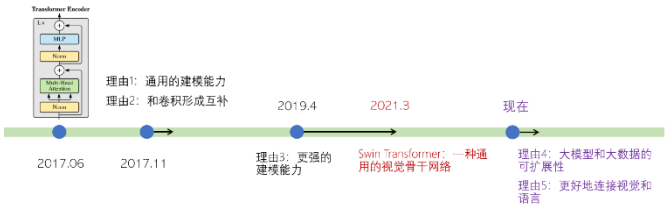


图 3　过去 4 年学界不断挖掘出的 Transformer 建模的五个优点

2.1. 通用的建模能力

Transformer 的通用建模能力来自于两个方面：一方面 Transformer 可以看作是一种图建模方法。图是全连接的，节点之间的关系通过数据驱动的方式来学习。由于任意概念（无论具体或抽象）可以用图中的节点来表示，且概念之间的关系可以用图上的边来刻画，因此 Transformer 建模具有很强的通用性。

另一方面，Transformer 通过验证的哲学来建立图节点之间的关系，具有较好的通用性：无论节点多么异构，它们之间的关系都可以通过投影到一个可以比较的空间里面计算相似度来建立。如下图右所示，节点可以是不同尺度的图像块，也可以是“运动员”的文本输入，Transformer 均可以刻画这些异构的节点之间的关系。

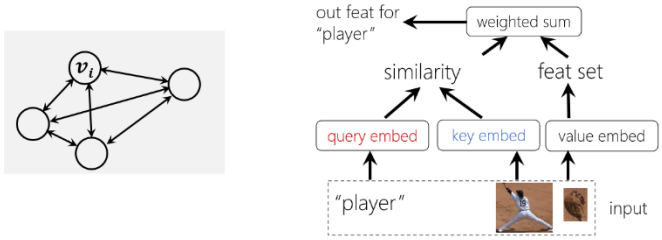


图 4　促成 Transformer 通用建模能力的两大原因：图建模（左）和验证哲学（右）

正是因为具备这样的通用建模能力，Transformer 中的注意力单元能被应用到各种各样的视觉任务中。具体而言，计算机视觉处理的对象主要涉及两个层次的基本元素：像素和物体。而计算机视觉所涉及到的任务主要就囊括了这些基本元素之间的关系，包括像素-像素，物体-像素和物体-物体的关系建模。此前，前两种关系

建模分别主要由卷积和 RoIAlign 来实现的，最后一种关系通常没有很好的建模方法。但是，Transformer 中的注意力单元由于其通用的建模能力，能被应用到所有这些基本关系建模中。近些年，在这个领域中已经出现了很多代表性的工作，例如：（1）非局部网络[17]。王小龙等人将注意力单元用于建模像素-像素的关系，并证明能帮助视频动作分类和物体检测等任务。元玉慧等人将其应用于语义分割问题，也取得了显著的性能提[18]。（2）物体关系网络[19]。注意力单元用于物体检测中的物体关系建模，这一模块也被广泛应用于视频物体分析中[20, 21, 22]。（3）物体和像素的关系建模，典型的工作包括 DETR[23]，LearnRegionFeat[24]，以及 RelationNet++ [25]等。

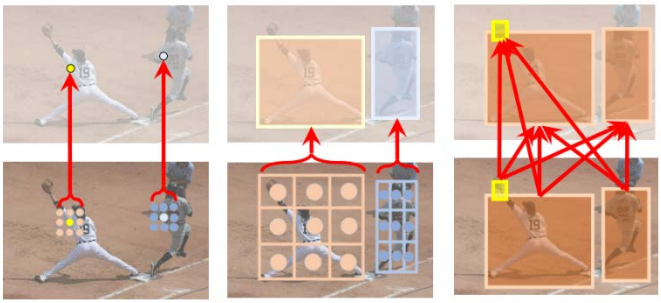


图 5　Transformer 能被应用于各种视觉基本元素之间的关系建模，包括像素-像素（左），物体-像素（中），物体-物体（右）

2.2. 和卷积形成互补

卷积是一种局部操作，一个卷积层通常只会建模邻域像素之间的关系。Transformer 是全局操作，一个 Transformer 层能建模所有像素之间的关系，它们能很好的互补。最早将这种互补性联系起来的是非局部网络[17]，在这个工作中，少量 Transformer 自注意单元被插入原始网络的几个地方，作为卷积网络的补充，并被证明在物体检测、语义分割和视频动作识别等问题中广泛有效。

此后，也有工作发现非局部网络在视觉中很难真正学到像素和像素之间的二阶关系[26]，为此，有研究员提出了一些针对这一模型的改进，例如解耦非局部网络[27]。

2.3. 更强的建模能力

卷积可以看作是一种模板匹配，图像中不同位置采用相同的模板进行滤波。Transformer 中的注意力单元则是一种自适应滤波，模板权重由两个像素的可组合性来决定，这种自适应计算模块具有更强的建模能力。

最早将 Transformer 这样一种自适应计算模块应用于视觉骨干网络建模的方法是局部关系网络 LR-Net[28]和 SASA[29]，它们都将自注意的计算限制在一个局部的滑动窗口内，在相同理论计算复杂度情况下取得了相比于 ResNet 更好的性能。然而，虽然理论上与 ResNet 的计算复杂度相同，但在实际使用中却要慢得多。一个主要原因是，不同的查询（query）使用不同的关键字（key）集合，如下图左所示，这对内存访问不太友好。

Swin Transformer 提出了一种新的局部窗口设计，称为移位窗口（shifted windows）。这一局部窗口方法将图像划分成不重叠的窗口，这样在同一个窗口内部，不同查询使用的关键字集合将是相同的，从而拥有更好的实际计算速度。在下一层中，窗口的配置会往右下移动半个窗口，从而构造了前一层中不同窗口像素间的联系。

2.4. 对大模型和大数据的可扩展性

在 NLP 领域，Transformer 模型在大模型和大数据方面展示了强大的可扩展性。下图中，蓝色曲线显示近年来 NLP 的模型大小迅速增加。大家都见证了大模型的惊人能力，例如微软的 Turing 模型、谷歌的 T5 模型以及 OpenAI 的 GPT-3 模型。

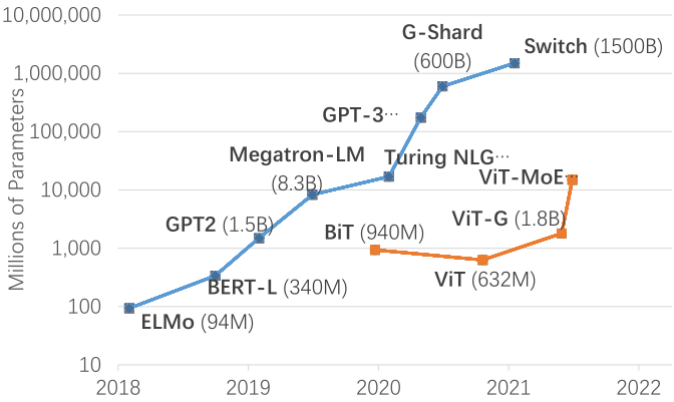


图 6　NLP 领域和计算机视觉领域模型大小的变迁

视觉 Transformer 的出现，也为视觉模型的扩大提供了重要的基础，目前最大的视觉模型是谷歌的 150 亿参数 ViT-MoE 模型[30]，这些大模型在 ImageNet-1K

分类上刷新了新的记录。

2.5. 更好的连接视觉和语言

在以前的视觉问题中，科研人员通常只会处理几十类或几百类物体类别。例如 COCO 检测任务中包含了 80 个物体类别，而 ADE20K 语义分割任务包含了 150 个类别。视觉 Transformer 模型的发明和发展，使视觉领域和 NLP 领域的模型趋同，这有利于联合视觉和 NLP 建模，从而将视觉任务与其所有概念联系起来。这方面的先驱性工作主要有 OpenAI 的 CLIP[31]和 DALL-E 模型[32]。

考虑这诸多优点，相信视觉 Transformer 将开启计算机视觉建模的新时代，我们也期待学界和业界能共同努力，进一步挖掘和探索这一新的建模方法给视觉领域带来的机遇和挑战。

责任编委 **魏秀参**

参考文献

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021.

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. ICCV 2021.

[3] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, Han Hu. Video Swin Transformer. Tech report 2021.

[4] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, Han Hu. Self-Supervised Learning with Swin Transformers. Tech report 2021.

[5] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, Jianfeng Gao. Efficient Self-supervised Vision Transformers for Representation Learning. Tech report 2021.

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, Manning Wang. Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation. Tech report 2021.

[7] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou. Training data-efficient image transformers & distillation through attention. Tech report 2021.

[8] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, Luc Van Gool. LocalViT: Bringing Locality to Vision Transformers. Tech report 2021.

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, Chunhua Shen. Twins: Revisiting the Design of Spatial Attention in Vision Transformers. Tech report 2021.

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. ICCV 2021.

[11] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, Shuicheng Yan. Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. Tech report 2021.

[12] Pengchuan Zhang, Xiyang Dai, Jianwei Yang, Bin Xiao, Lu Yuan, Lei Zhang, Jianfeng Gao. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. Tech report 2021.

[13] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. ICCV 2021.

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, Baining Guo. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. Tech report 2021.

[15] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao. Focal Self-attention for Local-Global Interactions in Vision Transformers. Tech report 2021.

[16] Zilong Huang, Youcheng Ben, Guozhong Luo, Pei Cheng, Gang Yu, Bin Fu. Shuffle Transformer: Rethinking Spatial Shuffle for Vision Transformer. Tech report 2021.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He. Non-local Neural Networks. CVPR 2018.
[18] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, Jingdong Wang. OCNet: Object Context for Semantic Segmentation. IJCV 2021.
[19] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, Yichen Wei. Relation Networks for Object Detection. CVPR 2018.
[20] Jiarui Xu, Yue Cao, Zheng Zhang, Han Hu. Spatial-Temporal Relation Networks for Multi-Object Tracking. ICCV 2019.
[21] Yihong Chen, Yue Cao, Han Hu, Liwei Wang. Memory Enhanced Global-Local Aggregation for Video Object Detection. CVPR 2020.
[22] Jiajun Deng, Yingwei Pan, Ting Yao, Wengang Zhou, Houqiang Li, and Tao Mei. Relation distillation networks for video object detection. ICCV 2019.
[23] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko. End-to-End Object Detection with Transformers. ECCV 2020.
[24] Jiayuan Gu, Han Hu, Liwei Wang, Yichen Wei, Jifeng Dai. Learning Region Features for Object Detection. ECCV 2018.
[25] Cheng Chi, Fangyun Wei, Han Hu. RelationNet++: Bridging Visual Representations for Object Detection via Transformer Decoder. NeurIPS 2020.
[26] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, Han Hu. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond. ICCV workshop 2019.
[27] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, Han Hu. Disentangled Non-Local Neural Networks. ECCV 2020.
[28] Han Hu, Zheng Zhang, Zhenda Xie, Stephen Lin. Local Relation Networks for Image Recognition. ICCV 2019.
[29] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, Jonathon Shlens. Stand-Alone Self-Attention in Vision Models. NeurIPS 2019.
[30] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, Neil Houlsby. Scaling Vision with Sparse Mixture of Experts. Tech report 2021.
[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. Learning Transferable Visual Models from Natural Language Supervision. Tech report 2021.
[32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. Zero-Shot Text-to-Image Generation. Tech report 2021.

胡瀚

微软亚洲研究院高级研究员。主要研究方向是：视觉表征学习以及视觉-语言联合表征学习。
Email: hanhu@microsoft.com

专题综述

全局和局部运动估计研究与展望

北京工业大学 毋立芳 电子科技大学 刘帅成 北京工业大学 相叶

一、引言

近年来自媒体、视频等可视媒体数据的爆炸式增长，对高质量的视频获取和智能化的视频分析提出了更高需求。运动是视频中的一个重要特征，它是获取视频的重要属性，也是视频对象行为的动态表达，在视觉关系表达、行为理解、视频稳像、视频对齐等应用中非常重要。

视频中的运动主要包括两类，一类是由相机运动导致的全局运动，如体育视频中的镜头运动、监控视频中外力影响导致的镜头晃动或者手持设备拍摄视频中的抖动等等；另一类是视频中的对象运动，称为局部运动，即视频中的运动主体发出的运动。另外还有一些与以上两类运动无关的区域如体育视频中的记分牌区域、LOGO 区域等。实际应用中，全局运动和局部运动可能存在关联性，如体育视频中的特定事件由特定运动员站位和动作组成，并且常用特定类型的相机表现手法，因而二者之间有一定关联。而在有些视频如自媒体视频或监控视频中，相机运动很多时候是干扰，不是关注重点。显然，有效估计全局和局部运动对于上述应用都非常重要。

目前大多数的运动估计方法如 PWC-Net[1]、循环全对场变换（RAFT）[2]等都是估计运动光流场，它包括全局运动和局部运动，我们称为混合运动。显然这种混合运动不能有效表达视频中的对象运动和对象行为，如图 1 所示。

除了光流场估计方法，也有一类专门估计全局运动的方法。包括传统方法[3]和基于深度学习的方法如深度单应性估计方法[4]、无监督深度学习方法[5] 、无监督深度单应性估计方法[6]等。基于深度学习的方法估计精度较高。然而这类方法无法得到视频中的局部运动。因此，有必要研究全局和局部运动估计方法。

图 1　视频图像和光流场，不同的颜色（灰度）代表不同的运动方向（幅度），不同对象颜色和亮度不同，说明其运动关联性较小。左图：篮球视频中的光流场包含全局运动、局部运动的混合运动以及静态区域；右图：相机静止，光流场表达局部运动

二、全局和局部运动分析

局部运动是视频中所有对象各自运动的总体表达，通常与对象行为或对象关系直接相关。然而作为不同的运动主体，大多数情况下，不同对象运动关联性较小，如图 1 所示。因此不同对象的运动幅度和运动方向不同，局部运动不具有移不变特性。

全局运动由相机运动产生，图像中不同位置的全局运动均服从于相同的相机运动，因此从系统的角度，全局运动具有空间移不变特性，它是一种线性移不变系统，可以用统一的系统函数来表达。在较远的场景如体育视频、监控视频中，场景较接近于一个平面，相机运动可

以用参数化模型来表达。结合全局运动的线性移不变特性，图像中全部像素点的坐标值变化均服从于统一的参数化全局运动模型，因此可以由局部区域的全局运动点来估计适用于整幅图像的全局运动参数。

相机运动包括平移(Translation)、旋转(Rotation)、缩放（Zoom in（out））、水平摇动（Pan）、垂直摇动（Tilt）。综合相应相机运动的参数化表达[8]，可以得到公式（1）。

$$\begin{cases} x' = (x+a) + (cx+d) + gx^2 + x + (\cos\theta x + \sin\theta y) \\ y' = (y+b) + (ey+f) + y + hy^2 + (-\sin\theta x + \cos\theta y) \end{cases} \quad (1)$$

进一步，可以表达为公式（2）。不同应用场景下，可以结合实际的相机运动进行简化。

$$\begin{cases} x' = m_2x^2 + m_1x + m_0 + m_3y \\ y' = n_2y^2 + n_1y + n_0 + n_3x \end{cases} \quad (2)$$

三、全局运动估计

全局运动估计的基本思路是由图像或者混合运动光流图估计相机运动参数或者全局运动图像，基本方法分为三类，分别是：传统方法、基于深度学习的方法和基于光流场分离的方法。

3.1. 传统方法

传统方法利用图像匹配技术，计算帧间变换模型，实现全局运动估计。常见的全局运动模型包括平移变换、仿射变换和单应性变换等。相较于平移、仿射，单应性变换具有更高的自由度。计算单应性变换通常需要检测和匹配特征点，比如 SIFT、SURF 等，然后通过鲁棒估计，如 RANSAC，剔除错误的匹配点，最后利用正确的匹配点拟合出单应性矩阵。

该类方法存在以下问题：（1）对于有重复纹理的场景 (比如很多建筑的窗户非常类似)，系统可能检测出很多特征点，但进行匹配时却不能有效一一对应；（2）弱纹理、无纹理。对于没有什么特征纹理的场景，系统很难在这些部分找出特征点；（3）大前景干扰。单应性变换只能拟合图像中的平面运动，非平面运动对应的特征点需要利用鲁棒估计加以排除。当图像中出现大前景干扰时，会对系统的鲁棒性造成很大挑战；（4）夜景、噪声干扰。在夜景、噪声干扰下，系统往往只能在一小块区域检测出特征点，然而用一小块区域来进行全局运动估计，效果往往不尽如人意。

3.2. 基于深度学习的方法

针对传统方法存在的问题，近年来一些研究者提出了深度学习的方法，实现更加鲁棒和准确的全局运动估计。DeTone 等人[4]通过给一个网络输入两张图像，可以直接得出单应性变换。该方法为有监督方法，模型的训练数据是人为对一张图像变形获得的，即随机产生一个单应性矩阵作为网络的监督，将该矩阵作用在任意一张图像上进行形变，形变前后的图像作为输入。然而，因为视差和运动物体的原因，真实世界不同图像之间除了角度变化还有内容上的差异。因此该方法在面对真实世界图像时效果不尽如人意。

Nguyen 等人[5]提出了一种无监督深度学习方法，通过优化图像对之间的损失，在真实数据上训练，从而克服了上述合成数据的局限。但该方法利用全图进行估算，没能有效排除图像中的运动区域和非平面区域。为了应对上述问题，Zhang 等人[6]采用无监督方法，对于输入的两张图像，提取深度特征的同时，估算一个 mask，其功能可类比为 NN RANSAC，从而剔除掉干扰区域，更鲁棒地回归单应性矩阵。

3.3. 基于光流场分离的方法

光流场是全局运动、局部运动以及场景无关区域的混合。根据线性移不变特性，可以从光流场中提取具有线性移不变性质的全局运动特征参数，再由这些特征估计出完整的全局运动。主要方法包括：

（a）基于统计分析的全局运动估计方法[7]

假设视频中不存在相机的扫换和旋转运动，则公式（2）可以简化为公式（3）。

$$\begin{aligned} x' &= m_1x + m_0 \\ y' &= n_1y + n_0 \end{aligned} \quad (3)$$

由公式（3）可以得到以下结论：（1）运动场包含水平和垂直方向两个通道，水平分量和垂直分量相互独立；（2）在 X（Y）方向的运动场分量中，运动幅度分布与

点的 Y（X）坐标无关，X（Y）坐标相同的点具有相同幅值。Y（X）坐标相同的点，运动幅度与 X（Y）坐标之间呈线性关系。

基于统计分析的全局运动估计算法基于以下常识—大多数情况下，视频边缘区域只包含全局运动。统计得到第 1 列（行）和最后一列（行）的水平（垂直）方向的运动幅值，以此为基础计算得到公式（3）中的参数，实现运动参数估计。统计分析法的优点是速度快，但是当视频边缘区域存在运动对象时，估计的全局运动模型参数会存在较大误差。

（b）基于迭代优化的全局运动估计[8]

体育视频转播过程中，常用到除旋转以外的相机运动，因此公式（2）简化为

$$\begin{cases} x' = m_2 x^2 + m_1 x + m_0 \\ y' = n_2 y^2 + n_1 y + n_0 \end{cases} \quad (4)$$

上式中 x 和 y 相互独立，因此可以分别估计其运动参数。将非全局运动视为异常点，利用光流中的全局点拟合相机运动模型，通过计算拟合误差逐步识别并舍弃数据空间中的异常点，提升全局运动估计结果的准确性，如图 2 所示。

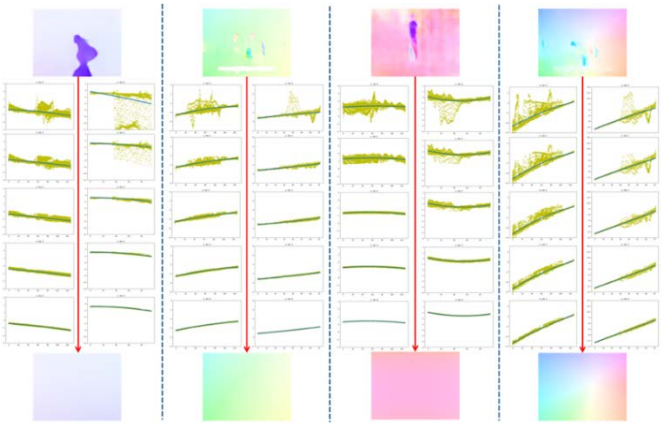


图 2　基于迭代优化的全局运动估计示例

（c）GLM-Net

迭代优化方法中 x 和 y 方向的运动估计相互独立，当图像中存在旋转时，该方法性能下降。针对这一问题，Yang 等人[9]提出了 GLM-Net，可以同时估计全局和局部运动的深度框架，如图 3 所示。设计 Mask Auto-encoder 框架实现全局运动估计的训练。将光流平铺为一维向量输入到网络中，考虑到相机模型表达最全参数为 8 个，设计编码网络最小降维到 8。其次通过解码网络将该向量解码为完整的全局运动。训练过程中，由于缺少完整的真实全局运动作为监督信号，因此，将光流中局部点的位置作为 mask 屏蔽掉，利用剩余的全局运动对网络输出进行约束。在验证阶段，网络无需监督即可从光流中估计完整的全局运动。

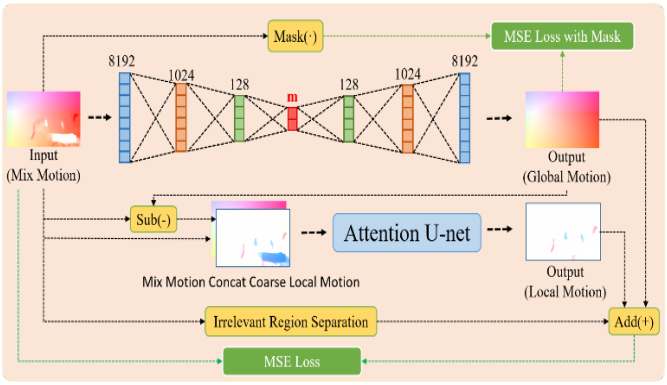


图 3　GLM-Net 网络框架

四、局部运动估计

最简单的局部运动估计方法就是由混合运动光流减去全局运动，但是当视频中存在运动无关区域如记分牌区域时，则该方法会引入运动无关区域而失效。考虑到这类运动无关区域多为静止区域，Wu 等人[7]提出了一种基于时空域阈值的局部运动估计算法。综合考虑并抑制连续 T 帧运动幅度低于设定阈值的像素点，从而有效去除场景无关区域。该方法的缺点在于需要连续 T 帧光流参与计算，无法利用单帧光流实现局部运动估计。GLM-Net（如图 3 所示）框架中，采用 Attention U-net 作为局部运动估计网络，将粗糙的局部运动和原始光流场拼接后作为输入，自动学习网络去除运动无关区域，输出优化后的局部运动区域。

五、应用

本节介绍全局和局部运动估计应用，图像数据库包括图像对齐数据库 DHE[6]和篮球比赛群体行为识别数据库 NCAA[10]以及个体行为识别数据库 UCF-101[11]。

5.1. 图像对齐

用估计得到的全局运动对第一张图像进行变换并与第二张图像进行对齐。DHE 数据库中，利用对应的两

张图像中人工标注的匹配全局点与图像对齐后的实际位置计算误差，评价图像对齐效果。对比实验结果如表 1 所示。可以看出 GLM-Net 能够得到与已有全局运动估计方法可比的结果。

表 1　不同方法的匹配点平均误差

	RE	LL	SF	LF
基于监督的方法	7.12	6.86	7.83	4.46
基于非监督的方法	1.88	2.27	1.93	1.97
SIFT + RANSAC	1.72	4.97	1.82	1.84
SIFT + MAGSAC	**1.71**	4.91	1.88	1.79
ORB + RANSAC	1.85	2.56	2.00	2.29
ORB + MAGSAC	2.02	2.78	1.92	2.25
LIFT + RANSAC	1.76	2.14	1.82	1.92
LIFT + MAGSAC	1.73	2.10	1.79	1.79
SOSNet + RANSAC	1.72	4.58	1.84	1.83
SOSNet + MAGSAC	1.73	4.39	1.76	**1.72**
CAU	1.81	**1.94**	**1.75**	1.77
GLM-Net	1.81	1.95	1.97	2.07

5.2. 相机运动估计

基于光流场分离的全局运动估计算法给出全局运动的可视化结果，进一步，由全局运动估计相机运动，不同算法的结果对比如图 4 所示，a 到 f 列分别为原始图像、原始光流、基于统计分析方法的全局运动估计结果、基于 RANSAC 方法的结果、基于迭代优化方法的结果以及基于 GLM-Net 方法的结果，每张图对应的两个数据是由全局运动估计的相机平移和缩放运动。可以看出，当视频边缘区域存在局部运动时，统计分析方法和 RANSAC 算法均存在较大误差。前者由于采用了边缘像素点，后者基于随机采样的点进行全局运动拟合，因此当局部运动在光流中占比较高时存在较大误差。第三行图像底部区域有运动无关区域，统计分析方法、RANSAC 算法和迭代优化算法的结果均不理想。由于该算法将底部垂直方向的局部运动拟合为全局运动，因此估计的全局运动中保留了部分局部运动。GLM-Net 算法在上述情况下均得到较好结果。

5.3. 行为识别

分别以原始混合光流和不同方法估计得到的局部运动作为输入，利用 3D 卷积网络（C3D、R3D、P3D、I3D）提取局部运动的时空特征进行行为识别。对比实验结果如表 2 所示。可以看出，基于局部运动的行为识别结果优于基于混合运动的结果，基于 GLM-Net 的结果略好于基于迭代优化的全局运动估计和基于时空域阈值的局部运动估计的结果。

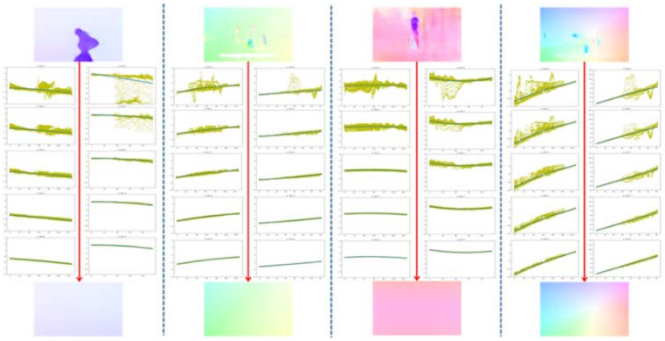


图 4　不同方法估计的全局运动和相机运动参数对比。(a) 原始图像（b）原始光流（c）基于统计分析方法估计的全局运动（d）基于 RANSAC 方法估计的全局运动（e）基于迭代优化方法估计的全局运动（f）基于 GLM-Net 方法估计的全局运动

表 2　行为识别结果对比

	UCF-101			NCAA		
	混合光流	迭代优化+时空域阈值	GLM-Net	混合光流	迭代优化+时空域阈值	GLM-Net
C3D	0.631	0.654	**0.684**	0.650	0.688	**0.690**
R3D	0.746	0.763	**0.775**	0.652	0.693	**0.702**
P3D	0.808	0.825	**0.839**	0.668	0.701	**0.711**
I3D	0.823	0.843	**0.859**	0.675	0.722	**0.731**

六、总结与展望

本文介绍了全局和局部运动估计方法及其应用，通过对混合运动光流场进行分离，能获得准确的全局运动和局部运动。实验结果表明，有效的全局和局部运动估计方法对于估计相机运动、图像对齐以及提升行为识别性能都有很大帮助。目前的研究还比较初步，后续有诸多改进点和探索点：在目前的光流分离中，如何有效的引入全局运动的物理意义、如何由相邻两帧图像直接估计全局和局部运动、如何利用相机参数对运动估计进行强约束、如何运用场景的深度信息进行引导等，都值得进一步研究。

责任编委　储珺

参考文献

[1] Sun D, Yang X, Liu M, Kautz J, Ieee. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)2018. p. 8934-43.

[2] Teed Z, Deng J. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow; proceedings of the Computer Vision – ECCV 2020, pp. 402-419.

[3] Harltey A and Zisserman A. Multiple view geometry in computer vision (2. ed.). Cambridge University Press.2006.

[4] DeTone D, Malisiewicz T, Rabinovich A. Deep Image Homography Estimation [J]. arXiv e-prints, 2016, arXiv:1606.03798.

[5] Nguyen T, Chen S, Shivakumar SS, Taylor C, Kumar V. Unsupervised Deep Homography: A Fast and Robust Homography Estimation Model [J]. IEEE Robotics and Automation Letters, 2018, 3(3): 2346-53.

[6] Zhang J, Wang C, Liu S, Jia L, Ye N, Wang J, Zhou J, Sun J. Content-Aware Unsupervised Deep Homography Estimation; proceedings of the Computer Vision – ECCV 2020, pp. 653-669.
数据库链接：https://github.com/JirongZhang/DeepHomography

[7] Wu L, Yang Z, Wang Q, Jian M, Zhao B, Yan J, Chen C. Fusing motion patterns and key visual information for semantic event recognition in basketball videos. Neurocomputing. 2020;413:217-29.

[8] Wu L, Yang Z, Jian M, Shen J, Yang Y, Lang X. Global motion estimation with iterative optimization-based independent univariate model for action recognition. Pattern Recognition [J]. 2021;116:107925. 代码链接：https://github.com/BJUT-VIP/Global-Motion-Estimation-with-iterative-optimization-based-Independent-Univariate-Model

[9] Yang Y, Xiang Y, Liu S, Wu L, Zhao B, Zeng B. GLM-Net: Global and Local Motion Estimation via Task-Oriented Encoder-Decoder Structure. ACM Multimedia (MM). 2021. 代码链接：https://github.com/BJUT-VIP/GLM-Net-Global-and-Local-Motion-Estimation-via-Task-Oriented-Encoder-Decoder-Structure

[10] Ramanathan V, Huang J, Abu-El-Haija S, Gorban A, Murphy K, Fei-Fei L. Detecting events and key actors in multi-person videos; proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), F 27-30 June 2016, 2016 [C]. 数据库链接：https://www.kaggle.com/ncaa/ncaa-basketball

[11] Soomro K, Roshan Zamir A, Shah M. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild [J]. arXiv e-prints, 2012, arXiv:1212.0402. 数据库链接：https://www.kaggle.com/pevogam/ucf101

毋立芳

北京工业大学信息学部教授，研究方向：视觉内容理解、智能 3D 打印、社会媒体计算。
Email: lfwu@bjut.edu.cn

刘帅成

电子科技大学信息与通信工程学院副教授，研究方向：图像/视频处理，底层计算机视觉。
Email: liushuaicheng@uestc.edu.cn

相叶

北京工业大学讲师，研究方向：视频群体行为识别，视频分析与处理等。
Email: xiangye@bjut.edu.cn

热点追踪

基于运动知识的视觉 SLAM 回环检测

中科院自动化研究所 刘秉熙 唐付林 傅禹杰 吴毅红

一、摘要

SLAM 系统在对未知环境的长期探索后，不可避免地产生轨迹预估误差和建图误差。视觉回环检测是对这一问题的公认解决方案，可以理解为一个在线的图像检索问题，要求实时、鲁棒的匹配当前地点与先前参观过的地点。基于局部特征的聚类技术广泛应用于回环检测，但不能很好地在移动平台上同时满足低时耗和高准确率。本文提出基于运动知识的视觉 SLAM 回环检测算法。这里的运动知识包括连续运动模型、基于网格的运动统计和运动状态区分。更进一步我们设计了一种灵活且有效的决策来决定局部特征和全局特征的使用。相关成果被 ICRA 2021 录取为口头报告。

二、引言

SLAM 系统在对未知环境的长期探索后，不可避免地产生轨迹预估误差和建图误差[1,2]。视觉回环检测是对这一问题的公认解决方案，可以理解为一个在线的图像检索问题，要求实时、鲁棒的匹配当前地点与先前参观过的地点，如图 1 所示。人工设计的全局特征计算较为快速，但易受光照、视角变化的影响。人工设计的局部特征鲁棒能解决视角问题，但计算比较耗时。局部特征的聚类技术被提出，其中基于无监督训练的词袋模型广泛应用于回环检测[1,2,3]。随着深度学习的发展，卷积神经网络在图像表达取得惊人的表现，同时逐渐被尝试应

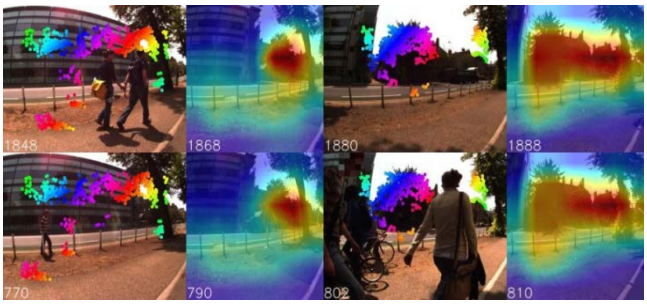


图 1　回环检测实例

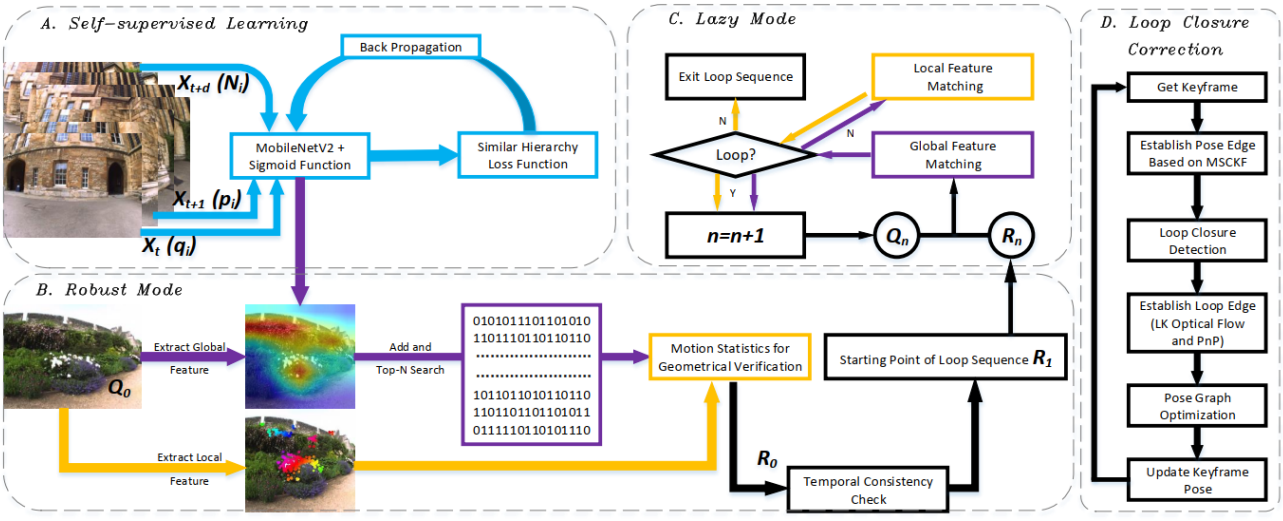


图 2　回环检测整体框架

用于位置识别和回环检测[4]。但最新提出的基于CNN的回环检测方法既没有考虑在移动平台上的实时运行表现，也没有充分融合运动学知识。因此，该工作着重研究一个与运动学知识紧密相关的 SLAM 回环检测和位姿优化系统，如图 2 所示。

三、正文

首先，一种基于连续运动模型的自监督标签方法被提出，即固定某个时间戳的图像为当前帧，时间序列上越是靠近当前帧的图像应该是更加相似的，而与当前帧相隔时间越长的图像相似度越低。被训练的轻量化网络用于提取图像的全局特征，两个全局特征之间可以快速计算汉明距离且准确地度量场景相似性。

仅依赖全局特征的检索是不鲁棒的且无法计算位姿，所以回环候选帧被提取局部特征并对这些特征进行匹配。在本文的研究工作中，综合评价了多个技术方案，选取了基于网格的运动统计的局部特征匹配方法作为回环检测系统的几何一致性检验模块，高效地解决了视角变化和遮挡问题。

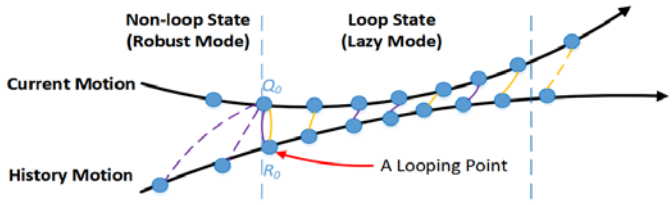


图 3　回环的抽象蓝图

最后，如图 2 所示，我们通过区分运动状态，设计了一个充分利用线性存储结构和有机融合全局和局部特征的检测策略。当一个运动系统到达回环点时，它在将接下来的一段时间处于一个回环路径。因此，我们区分运动状态为非回环状态和回环状态，如图 3 所示。假设查询图像Q_0检测到R_0，则后续帧Q_i会检测到R_i。被区分的两种运动状态分别对应回环检测系统中的两种模式：鲁棒模式和偷懒模式。鲁棒模式下，我们利用了全局特征检索、局部特征验证和时间一致性验证；偷懒模式下则是两种特征自适应交替使用，目的是提高检测速度的同时可以适应尺度或视角变化较大的场景。

表 1　位姿优化过程中的各项实验数据

Stages	Outdoor1	Outdoor2
Number of Keyframes	1101	432
Total Optimization Time (ms)	98.14	33.24
Mean Optimization Time (ms/keyframe)	0.029	0.077
Reprojection Error (pixel)	1.53	1.83

表 2　100%准确率下不同算法的召回率

	City Centre	New College	KITTI 00	KITTI 05
FAP-MAP 2.0[1]	40.11	52.63	61.22	48.51
DLoopDetector[2]	30.59	47.56	72.43	51.97
An et al.[3]	66.48	76.74	91.23	85.15
Tsintotas et al.[4]	52.44	16.30	**93.18**	**94.20**
Proposed	**86.01**	**91.21**	93.02	92.53

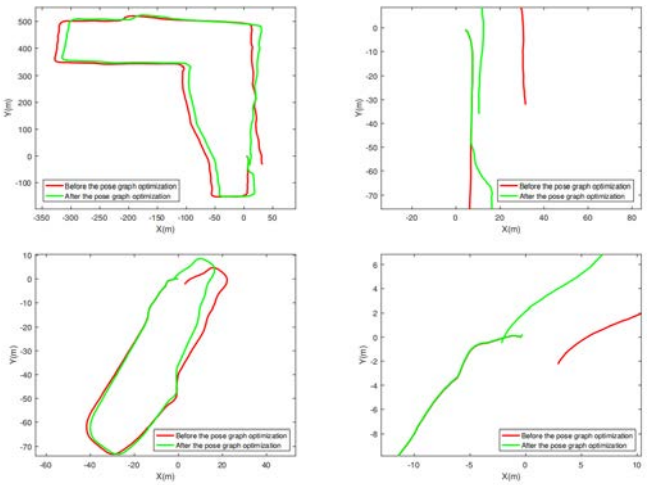


图 4　真实场景下的回环检测及误差矫正

利用上述回环检测系统和课题组的 VIO 系统，进一步设计了一个位姿优化模块用于纠正累计误差。我们提出的系统在多个公开数据集和真实场景数据下进行了大量实验，并与先进方法进行了结果对比。表 1 展示了被量化的平均优化时间和重投影误差。图 4 展示了真实场景下的位姿图优化前后的轨迹。表 2 展示我们的算法和其他先进算法在公开数据集下结果对比，评价指标是 100%准确率下的召回率。我们提出的算法在 New College 上比结果最好的算法要高出 14.47%。更重要的是，我们的结果在多个数据集下是比较稳定的。

责任编委　崔海楠

参考文献

[1] Cummins M, Newman P. Appearance-only SLAM at large scale with FAB-MAP 2.0[J]. The International Journal of Robotics Research, 2011, 30(9): 1100-1123.

[2] Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences[J]. IEEE Transactions on Robotics, 2012, 28(5): 1188-1197.

[3] Yue H, Miao J, Yu Y, et al. Robust Loop Closure Detection based on Bag of SuperPoints and Graph Verification[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2019: 3787-3793.

[4] Tsintotas K A, Bampis L, Gasteratos A. Probabilistic appearance-based place recognition through bag of tracked words[J]. IEEE Robotics and Automation Letters, 2019, 4(2): 1737-1744.

刘秉熙

中科院自动化研究所硕士生。主要研究方向为视觉定位。
Email: bingxi.liu@nlpr.ia.ac.cn

唐付林

中科院自动化研究所助理研究员。主要研究方向为 SLAM。
Email: fulin.tang@nlpr.ia.ac.cn

傅禹杰

中科院自动化研究所博士生。主要研究方向为图像匹配。
Email: yujie.fu@nlpr.ia.ac.cn

吴毅红

中科院自动化研究所研究员。主要研究方向为相机定位与标定、三维重建、SLAM 等。
Email: yhwu@nlpr.ia.ac.cn

顶会观察

CVPR 2021

北京邮电大学 邓伟洪

国际计算机视觉与模式识别会议（IEEE Conference on Computer Vision and Pattern Recognition，CVPR）是计算机视觉及模式识别的顶级学术会议，与 ICCV、ECCV 并称为计算机视觉领域三大顶会。CVPR 有着非常高的学术影响力，在中国计算机学会推荐国际学术会议中被评为人工智能领域的 A 类会议；在 Google Scholar 发布的学术指标中，其 H 指数雄踞人工智能领域的榜首。引人注目的是，今年大会组委会有不少华人面孔：中国科学院院士谭铁牛担任 General Chair，上海科技大学信息科学与技术学院教授虞晶怡、肯塔基大学计算机系终身教授杨睿担任 Program Chair，中山大学智能工程学院副教授梁小丹担任 Tutorials Chair 等。

一、 国际计算机视觉与模式识别会议的亮点

CVPR 2021 于美东时间 2021 年 6 月 19 日至 25 日在线上举行。由于疫情原因，会议主办方建立了虚拟会议的网站，以供参会人员进行展示及技术交流。作者需要为每篇论文准备一个五分钟的预先录制的视频和海报的 PDF 文件来演示工作，参会者可以按需查看演示文稿和视频。同时，会议为每篇论文安排了指定的线上交流时间，允许作者与感兴趣的参会者通过文本聊天或线上会议的方式进行交流。CVPR 研讨会以及教程将通过直播视频进行，主持人和参与者之间进行现场问答。会议还包括具有视频和文本聊天元素的多个在线网络活动，为参会者提供了自由、便利的会议环境，使每位参会获得了良好的参会体验及交流经历。

除了后面将介绍的最佳（学生）论文奖外，大会还颁发了两项传统大奖。Longuet-Higgins Prize 以认知科学家 H. Christopher Longuet-Higgins 的名字命名，表彰十年前对计算机视觉研究产生重大影响的 CVPR 论文。今年的获奖论文是来自微软的 Real-time human pose recognition in parts from single depth image（从单一深度图像中实时识别人体姿势）和来自石溪大学的 Baby talk: Understanding and generating simple image descriptions（婴儿谈话：理解和生成简单的图像描述）。Young Researcher Awards 旨在表彰对计算机视觉做出杰出研究贡献的年轻研究人员，今年获奖者是 FAIR 的 Georgia Gkioxari 和 MIT 的 Phillip Isola。

今年大会的亮点是首次颁发的 Thomas S. Huang 纪念奖。该奖项为了缅怀一代 CV 宗师、华人计算机视觉泰斗 Thomas S. Huang（黄煦涛），由 PAMITC 奖励委员会选出。今年首届获奖者是 MIT 电子电气工程与计算机科学教授 Antonio Torralba。Torralba 的研究领域包括场景理解和上下文驱动的目标识别、多感官知觉整合、数据集构建以及神经网络表征的可视化和解释。

二、 论文录用情况

CVPR 2021 总共收到了 7,039 篇有效投稿，其中 1,661 篇论文被接收，接收率约为 23.5%，相比 CVPR 2020 论文接受率略有回升。其中，有 295 篇论文入选 oral presentation，oral 率约为 4.1%，低于去年的 5.7%。同时，从大会公布的数据来看，今年大会接收到的注册及有效投稿数量都有显著的提高。CVPR 2021 会议涵盖的方向包括目标检测、行为识别、对抗攻击与防

御、生物特征、计算摄影、图像和视频检索、图像和视频合成、图像分类、姿态估计、无监督学习、视频理解、多模态等方向。在 CVPR 2021 接收的论文中，3D 视觉、计算摄影学、视频图像合成三个方向的论文数量最多，无监督、半监督、自监督三个关键词的出现次数相较 CVPR 2020 上升了 50%。在 CVPR 2021 最佳论文奖的 32 篇候选论文中，有华人参与的论文高达 18 篇，华人为一作的论文共有 16 篇，且其中 6 篇的一作为国内机构学者。本次 CVPR 2021 收录论文中，来自中国工业界的各大互联网企业获得了不俗的成绩。根据公开数据，商汤及联合实验室共 66 篇论文入选，腾讯 AI 实验室与优图团队共有 33 篇论文入选，华为诺亚方舟研究团队有 30 篇论文入选，旷视有 22 篇论文入选。这些被录用的论文在很多重要工业应用领域上取得了重大突破，包括模型压缩、网络架构搜索、语义理解、底层视觉、光流估计、无监督学习、人体姿态估计、目标检测等。此外，谷歌在本次会议表现依旧亮眼，共有 70 余篇论文入选， 其中华人为第一作者的论文共有 34 篇。

三、 主题演讲

为了方便深入交流，大会将受邀演讲者的研究领域大致分为三组：AI 伦理、计算机视觉中的机器学习、人类和机器人感知。

安第斯大学的 Pablo Arbelaez 博士带来了“人工智能促进全球健康”的演讲。 Google Ghana 的 John Quinn 博士的演讲从乌干达和加纳团队的工作角度，讨论了计算机视觉如何为健康、气候和粮食安全这些联合国关注的与持续发展相关的领域的进步做出贡献，以及所面临的困难和风险。麻省理工学院的 Catherine D’ Ignazio 教授认为伴随海量数据而来的不平等的生产条件、不对称的应用方法以及它们对个人和群体的不平等影响越来越难以被数据科学家和其他在工作中依赖数据的人忽视，并讨论了如何实现合乎道德和公平的数据使用。

麻省理工学院的 Constantinos Daskalakis 博士从优化、复杂性理论和拓扑方法等方面阐述了对均衡计算与多智能体学习的见解。纽约大学的 Meredith Whittaker 教授探讨了美国军方与计算纠缠的历史，并且其认为冷战思维的延续产生了“人工智能军备竞赛”，使美国认为必须赢得这场竞赛才能维持军事和经济霸权。南京大学的周志华教授简要介绍了学习理论研究的悠久历史和关于 Boosting 的争论，并揭示了在学习过程中最大化边际均值时最小化边际方差的重要性，以及其为强大学习算法的设计所带来的灵感。

加州理工学院 Katie Bouman 讲述了如何利用计算成像管道代替传统光学成像，以获取传统光学成像无法获取到的图像。加州大学圣克鲁兹分校的 Su-hua Wang 博士分享了人类婴儿如何在动态事件中学到物体的表示以及不同的模式，图宾根大学 Matthias Bethge 展示了为了让机器像人类一样看待事物所进行的工作，百度的 Liang Huang 介绍了机器在同声翻译领域近期的突破与进展。此外，一个名为“计算机视觉遇见安全”的附加小组会议探讨了计算机视觉安全基础对技术、市场和研究的相关影响。

四、会议获奖和热点论文、教程与竞赛

最佳论文奖评审委员会由 CVPR 领域的 9 名国际权威学者组成，包括来自中科院计算所视觉信息处理与学习组的陈熙霖教授。今年大会共评选出了 1 篇最佳论文，2 篇最佳学生论文，2 篇最佳论文提名，3 篇最佳学生论文提名。

最佳论文：GIRAFFE：Representing Scenes as Compositional Generative Neural Feature Fields 的主要贡献是将组合式三维场景表示纳入生成模型，使得图像合成更加可控。该模型将图像场景表示为生成神经特征场的组合，可以在无需任何额外的监督的情况下，从非结构化和非特定视角的图像集合中学习到如何将每个对象的形状和外观从背景中分离出来。通过将场景表示与神经渲染管道（Neural Rendering Pipeline）相结合获得一个快速而逼真的图像合成模型，能够分解单个物体，并允许其在场景中平移和旋转，还可以改变摄像机的姿势。

最佳学生论文： Task Programming: Learning Data Efficient Behavior Representation 提出一种用

于减少自动行为分析任务中数据集标注工作量的方法。该论文基于多任务自监督学习，提出了一种用于行为分析的有效轨迹嵌入方法—TREBA。利用该方法专家们可以通过“任务编程”过程来有效地设计任务，即使用程序编码将领域专家的知识结构化。通过交换数据注释时间来构造少量编程任务，可以有效减少领域专家的工作量。在行为神经科学领域的数据集上，小鼠和果蝇两个领域内三个数据集的测试评估表明，TREBA 使注释负担减少到原来的十分之一。

最佳论文提名：Exploring Simple Siamese Representation Learning 发现了简单的孪生网络可以学习有意义的表示，实验验证了：即使不使用 1）负样本对，2）大 batch，3）momentum 编码器中的任何一项，也能获得很好的自监督学习效果。Learning High Fideity Depths of Dressed Humans by Wathing Socail Media Dance Videos 研究了如何高效利用社交媒体中大量的舞蹈视频进行自监督学习：将预测的局部几何体从一幅图像在不同的时刻扭曲到另一幅图像，使得自监督学习对预测实现时间的一致性。

最佳学生论文提名：Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling 表明用少量稀疏采样片段的端到端学习会比使用从全长视频中密集提取的离线特征更加准确，验证了视觉模型训练样本中的“少即是多”原则。Binary TTC: A Temporal Geofence for Autonomous Navigation 通过一系列简单的二元分类来估计场景的相对深度 (Time-to-Contact)，对于视觉导航有重要意义。Real-Time High-Resolution Background Matting 提出了一种实时、高分辨率的背景更换技术，该技术可以在 GPU 上以 30fps 速度运行 4K 分辨率和以 60fps 的速度运行高清分辨率。

另外，来自国内的多篇论文也引起了广泛的讨论。北大的论文 Generalizing to the Open World: Deep Visual Odometry with Online Adaptation 提出了一种结合了深度学习和几何计算优点的在线自适应框架，使得深度视觉里程计网络能够以自监督的方式快速适应新的场景，在多个数据集上实现了更好的泛化性能与深度估计性能。中科院自动化所的论文 Information Bottleneck Disentanglement for Identity Swapping 提出了一种基于信息瓶颈解耦的高身份辨识度换脸方法，通过约束互信息、学习身份信息的最小充分统计量，将相互耦合的身份信息（identity）和感知信息（perception）显式地分流，生成了高身份可辨识度的图像，并根据对比学习的思想提出了一项度量身份可辨识度的统计评价指标，有力地推动了换脸技术的发展。中科院计算所的论文 FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation 针对图像描述生成任务，借鉴人类语言翻译中的“信-达-雅”分级评价思想，提出了以图像描述忠实性和充足性为核心的层次化评价指标，并通过构建视觉与语言跨模态场景图以对齐多粒度图文信息，将图像描述生成的评价问题形式化为多实例多模态场景图匹配问题，获得了与人类评价结果高度一致的图像描述评价系统。微软亚洲研究院的工作 DEKR：Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering 提出了一种基于密集关键点坐标回归的多人姿态检测模型，以解决在拥挤的人群的场景下，由于人群过于密集，重合程度太高，导致每个人的位置难以用人体检测框表示的问题，达到了目前自底向上姿态检测的最好结果。腾讯优图与南京理工大学大学合作论文 Consistent Instance False Positive Improves Fairness in Face Recognition 提出了一种基于误报率惩罚的损失函数，在无需人口统计学标注的情况下，通过增加实例误报率（FPR）的一致性来减轻人脸识别模型在基于不同属性划分的人脸组上的性能偏差，缓解了人脸识别中的不公平问题。华南理工的论文 Implicit Feature Alignment: Learn to Convert Text Recognizer to Text Spotter 提出了一种被称为隐式特征对齐的方法，该方法可以被集成到普通的文本识别器中，使其能够处理多行文本，在多个文档识别任务上获得了最先进的性能。

此外，CVPR 2021 包含精心组织举办的 83 个研讨会（Workshops）与 30 个讲习班（Tutorials），涵盖了可解释机器学习、细粒度视觉、视觉数据压缩、对抗鲁棒性、自监督学习、自动驾驶、视听场景理解、医学计算机视觉等多个领域。这些研讨会与讲习班为参会者提

供了优质的交流平台与学习资源，同时帮助参会者拓宽了学术视野。在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突出的成绩。由欧洲科学院外籍院士焦李成领衔的西电人工智能团队在洪水中高分辨率 UAS 图像的分类与语义分割、洪水中高分辨率 UAS 图像的视觉问答、无监督二类地物分类变化检测等赛事中，获得 4 项冠军的优异成绩。百度在此次参与的 7 项挑战赛中共获得 10 项冠军，涉及自动驾驶、人体解析、智慧城市、物体检测、图像修复增强、视频目标分割、视频理解等多个方向。

五、总结与展望

回顾近几年的 CVPR 获奖和热点论文，我们可以发现，CVPR 越来越青睐致力于解决真实场景下存在的视觉问题的方法或工作，包括对真实场景的建模、实现各类视觉子任务（分类、检测、语义分割）的统一融合，从而模拟人类视觉系统对真实物理世界的认知。随着相关算法与硬件计算能力的不断升级，3D 视觉算法效果得到大幅提升，三维几何重建更加精细，表面纹理重建更加清晰，正为我们带来更加逼真的视觉观感和数据生成效果。无监督学习和弱监督学习通过不使用标签或减少对标签数量、质量的要求来迅速降低深度模型对于数据的标注需求，大幅提高了数据的利用效率，正在由量变引发质变。随着视觉认知能力的提升，多模态融合也从感知视觉内容，逐渐拓展到学习物理关系，逻辑推断，因果分析等知识，正在从感知智能迈向认知智能。

责任编委 王金甲

邓伟洪

北京邮电大学“鸿雁人才”教授，教育部青年长江学者。研究方向为生物特征识别、可信人工智能、情感计算、多模态学习。曾入选北京市优秀博士学位论文、教育部新世纪优秀人才、北京市科技新星、Elsevier 中国高被引学者等。

Email: whdeng@bupt.edu.cn

中科院心理所王甦菁副研究员访谈

2021 年 9 月 5 日，《CCF-CV 专委简报》在线采访了中科院心理所博士生导师王甦菁副研究员。下面是采访实录。

王老师，您好！首先，请您分享一下您的个人学习和研究经历。

关于我对计算机的学习和研究，说来话长，记得我在 10 岁左右的时候偶然从一本书中了解了计算机和编程，那时就引起了我的兴趣。后来慢慢学习了 BASIC 语言。记得我上初中三年级的时候，数学老师在课堂上讲数列，我在下面在草稿本上用 BASIC 语言写相应的代码。等到我上高一的时候我才有了人生的第一台电脑，80286，640K 内存、单显、无硬盘，只有两个软盘驱动器。自此，我就开始沉迷于电脑编程。1995 年我参加高考时，一心只想报考一所有计算机系的学校，能够在有老师指导下学习计算机，但是因为当时我国对残疾人权益的保障不健全，没有高校愿意招收我，我只能在家读广播电视大学，而且电大没有计算机专业，我只能读财会大专。于是我一边读财会，一边在家继续自学计算机。当时有“计算机技术与软件专业技术资格（水平）考试”，我就报名参加了这个考试，在参加我们市科技委员会组织一次关于如何准备这种考试的培训上，有幸见到了我国第一代计算机教育专家，南京大学钱士钧教授。钱教授对我说：“我们系有个小伙子，搞得挺不错的，你以后可以关注一下他。”从那时起，我就一直关注这个小伙子，但感觉我和这个小伙子的差距越来越大。这个小伙子就是南京大学周志华教授。同时为了应考“计算机技术与软件专业技术资格（水平）考试”高级程序员级，我自己写了“汇编语言编译器”。1997 年由钱士钧教授推荐，我代表江苏省广播电视大学参加了“首届中国大学生电脑大赛”。在总决赛开幕式时，受到了时任国务院副总理邹家华的亲切接见。前几年和黄铁军教授交流时，才知道，黄铁军教授当年是在读博士，协助指导华中科技大学辩论队也参加了这个比赛，得了冠军。

1997 年电大毕业后，我还想通过读研究生来继续研究计算机，但是报考研究生需要具有学士学位。所以我又花了三年读了个成人本科，拿到了学士学位。然而考研对于我来说，还有一个更严峻的问题，那就是考试时间的问题，因为我的手写字很慢，我写字的速度相当于一般人写字速度的 1/8。所以我无法在规定的时间内答完题。这样考了四、五年都没有能达线。最后，2005 年去了吉林大学读了软件工程硕士（非学历教育）。2008 年，才正式进入吉林大学计算机科学与技术学院，攻读博士学位，师从著名计算机专家周春光教授，进行计算机视觉的研究。2012 年进入中国科学院心理研究所进行微表情分析相关研究。

您在微表情分析领域很有建树，能否介绍一下您在这些领域中最突出的几项研究成果？

其实我没有什么建树，只不过是在微表情分析领域

中比大家先走了几步。至少对于计算机视觉来说，数据是基础。当我进入中国科学院心理研究所做博士后，就在我的合作导师傅小兰研究员的指导下，建立诱发自然微表情的心理学范式，并在此基础上建立并公开发布了中国科学院微表情（CASME）系列数据库，为微表情识别算法的发展提供了数据基础。发现了微表情的稀疏表示有助于细微变化表达的现象，通过稀疏张量表示微表情的结构信息。把 Faster R-CNN 中在空间域上提出的锚点（anchor）的概念引入时间域，并使用一维扩张卷积(1D dilated convolution)实现了时间域上的锚点，从而实现一个微表情检测的网络。

您近年发表了多篇顶刊或顶会学术论文，而且您入选 2020 全球前 2%顶尖科学家“年度影响力”榜单，能跟大家分享一下您是如何做到持续产出高水平论文的么？您在学术影响力方面做了哪些努力呢？

“Research Interests”经常在国内被对应于“研究方向”。其实兴趣才是最好的老师，兴趣才是最本质的驱动力。我享受每一次成功地对审稿人的 rebuttal 和对代码的 debug。另外，我也重视学术交流，在交流的过程中，能够碰撞出思维的火花。我相信很多委员都能在咱们专委会的年会上看到我的身影。我也会把会议上听到的很多报告，无论是动作识别，还是人脸识别等，把这些研究方向的学术报告中讲的技术，有选择的放在微表情分析上试一试。

您获批了多项纵向课题，能跟大家分享一下您成功申请的经验和体会吗？

我做一个课题时，会发现一些新的问题，对这些问题梳理后，就可以形成下一个课题的申请书。前后两个课题之间有着一定的联系，前一个课题可以作为后一个课题的研究基础。这样申请可能会比较容易获批。另外，自己做的课题也要的当时流行的一些技术想结合。比如我 2013 年获批的第一个国家自然科学基金委面上项目名称叫“基于稀疏张量的微表情识别研究”，当时稀疏编码技术在计算机视觉中很热。2017 年获批的第二个国家自然科学基金委面上项目名称叫“基于深度学习的微表情检测和识别的研究”，这些年，深度学习技术很火。

可否请您谈一下在第三代人工智能时代，微表情识别将如何发展？面临哪些挑战？哪些研究方向会特别有价值呢？

随着人工智能逐渐融入人们的日常生活，以及深度学习可解释性的不断挖掘，图像采集设备精度的提升，微表情识别的研究也将逐渐从实验室走向现实应用，例如审讯场景中谎言检测的辅助信息，医疗关怀中对病人异常情绪的检测等。但是，由于微表情本身非常微小、短暂，同时微表情样本的诱发、采集和标注都十分困难，造成微表情的样本数目很少，并且目前已经公开的微表情数据库生态效度比较低，创建一个具有鲁棒性的智能微表情分析系统十分困难，也限制了基于深度学习的微表情分析算法的研究。

在未来的研究工作中，对微表情的生理生成机制的探索将有助于夯实微表情在实际场景应用的理论基础；半自动乃至全自助的微表情标注也是值得关注的工作，能够直接为大数据驱动的算法研究提供丰富样本；随着各种多模态信号采集设备在日常生活的普及，融合生理信号、语音信号等模态的多模态微表情识别分析将提升对个体异常情绪检测的灵敏度。

您被新华社称为“中国版霍金”，这背后有什么样的传奇与人生哲学呢？另外，您热衷公益，请问您是怎样兼顾科研和公益活动的呢？

有可能是我在中国科研工作者中，身体情况最像霍金教授的吧，我也说不清楚。但我知道科研的目的不是发论文，也不是拿项目，而是为了改善和提高人民生活质量。我既是一位人工智能科研工作者也是一名残疾人，首先就是想到，如何用人工智能技术来改善残疾人的生

活质量。我在 2020 年 CNCC 上组织举办了一公益性技术论坛“AI+辅具:让人工智能提高残障人士生活质量”。全国人大常委、中国残联吕世明副主席作为论坛的嘉宾莅临这个论坛，并充分地肯定了该论坛的意义。

请问您是如何管理和运作您的团队的？您是如何管理研究生的？您对他们的要求是什么？

我的实验室叫 MELAB (Micro-Expression laboratory)，其另一层意思也是“我的实验室”，希望团队中的每一位成员都有主人翁意识，希望能把实验室打造成温馨的家。我在尽我所能为学生提供学习和生活环境，像家长一样关心，像朋友一个沟通，亦师亦友亦兄弟，要和他们讲“哥们义气”。我也相信“不护犊子的老师不是好程序员”。我不敢说 MELAB 是心理所产出最多的实验室，但我敢说 MELAB 是心理所产出笑声最多的实验室，没有之一。

您能获得今天的成就，与您家人的支持一定是密不可分的，可否谈谈家人对您的支持？

当然我之所以有今天，和我的家人支持是密不可分的。大家都知道做科研都需要基金支持。就拿我的第一台 286 电脑来说，当时购买花了一万多元。在上个世纪九十年代，我不确定这是否相当一个青年基金的支持力度，然而我购买电脑这些钱全来自于我的家。不仅是我的父母，而且还有我的外公、外婆、姨夫、姨妈、舅舅、舅妈一直在滚动地资助我。在此，请允许我对他们表示感谢！

如果吐露研究工作者的心声，您最想说的是什么？

享受科研、享受生活!!! 最后我还想说的是“我长期招收从事微表情研究的博士后，欢迎各位委员推荐!!!”

责任编委 赵振兵 余烨

王甦菁

王甦菁，中国科学院心理研究所副研究员，博士生导师。2012 年 6 月博士毕业于吉林大学计算机科学与技术学院，2012 年 8 月至 2015 年 6 月在中国科学院心理研究所做博士后工作。2015 年 7 月加入中国科学院心理研究所。主要研究方向为模式识别与机器学习，特别是微表情识别。在国内外重要期刊和学术会议上发表五十余篇论文，包括 TPAMI、TIP、TNN、ECCV 等。2014 年起担任 Neurocomputing 期刊的 Associate Editor。CCF 杰出会员，IEEE 高级会员，中国计算机学会计算机视觉专业委员会委员，中国人工智能学会人工心理与人工情感专业委员会委员，中国图象图形学学会机器视觉专业委员会委员。主持国家自然科学基金面上项目 2 项，北京市自然科学基金面上项目 1 项，中国博士后基金 2 项。获 2018 年第八届吴文俊人工智能科学技术奖一等奖。入选 2020 全球前 2%顶尖科学家“年度影响力”榜单。被新华社称为“中国版霍金”。

委员好消息

✪ 2021 年 7 月 1 日，由 CCF-CV 专委会常务委员、北京大学**林宙辰**教授等合著的《机器学习中的加速一阶优化算法》一书由机械工业出版社出版，上市一周即荣登京东新书榜第一位。该书全面介绍机器学习加速算法以及最新进展，包括确定性和随机性的算法、同步和异步的算法，以求解带约束的问题和无约束的问题、凸问题和非凸问题，融合扎实工作、深刻见解和不同领域问题的细致分析,对算法思想进行了深入的解读，并对其收敛速度提供了详细的证明。

✪ 2021 年 7 月 27 日，黑龙江省教育厅发布了《关于表彰黑龙江省第十二届普通高等学校第三届中等职业学校教学名师奖获得者的决定》，CCF 专委会委员、哈尔滨工程大学**刘海波**教授获得省级教学名师奖。

责任编委 刘海波

基于骨架数据的人体动作识别开源代码

北京航空航天大学 张奇鹏 王田

基于骨架数据的人体动作识别是对序列数据进行处理，给出序列的动作类别。近年来，由于深度传感器的进步，学者开发了一系列公开的人体骨架数据集，旨在帮助基于骨架数据的人体动作识别和预测的技术发展，这对于人体动作的识别有着深远的研究意义和重大的战略价值。例如，在监控安保领域，可以利用人体动作识别技术，甄别出一些可能发生危险或者将要产生犯罪的动作，极大的保证公共场所的安全秩序。而对于视频处理领域，可以使用动作识别技术来对视频中的一些特定动作进行分类，从而达到对视频流检测的效果，极大的减少暴恐视频的传播。而在日益火爆的自动驾驶方面，人体动作识别技术更是能发挥巨大的作用。对于自动驾驶汽车来说，可以使用动作识别来分析道路上行人的走势以便判断他们下一时刻的位置，这样自动驾驶算法可以尽早的做出决策。

随着图神经网络的发展，研究人员开始将图结构引入到人体骨架数据上，有更多的研究开始关注人体骨架图的建模。本文将从 ST-GCN 开始，重点介绍几种关于人体骨架图的建模，以及 Transformer 和 GCN 相结合的一些研究成果。

1、ST-GCN

工作：该论文作为骨架动作识别领域应用图神经网络的开山之作，提出了一种分层结合 GCN 和 TCN 的模型。如图 1 所示，整体的网络结构，对视频利用 Open Pose 等算法进行姿态估计，并构造骨架序列的时空图。之后用多层的时空图卷积(Spatial Temporal Graph Convolutional, ST-GCN)，逐步在图上生成更高层次的特征图，最后用标准的 Softmax 分类器将其分类为相应的动作类别。

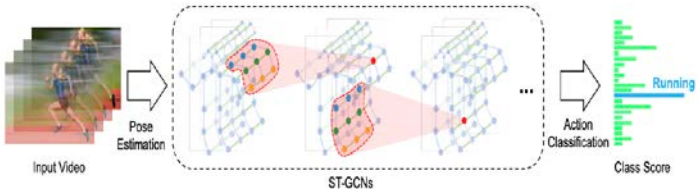


图 1 ST-GCN 结构

论文：https://arxiv.org/abs/1801.07455

代码：https://github.com/yysijie/st-gcn

2、DGNN

工作：该论文首次将人体骨架图考虑为有向图，如图 2 所示，关于两个点组成的边的方向，作者定义为离中心点近的点指向离中心点远的点。其模型结构类似于 ST-GCN，采用 9 层 DGN 和 TCN 来对人体骨架序列进行建模，逐步提取人体骨架图的信息，后续类似于 ST-GCN，采用全局池化和 Softmax 进行分类。其特点在于，DGN 每一层输入的是包含顶点和边的属性的图，而输出是更新后包含顶点和边属性的图。也就是说每经过一层，顶点和边的特征就得到更新，而不同于 GCN 只包含节点的信息。在最底层网络，每个顶点和边能接收到的属性就是邻近的边和点，所以只能表示局部特征。而随着训练的深入，在最顶层网络，相聚比较远的顶点

和边也能联系到一起，从而得到全局特征。

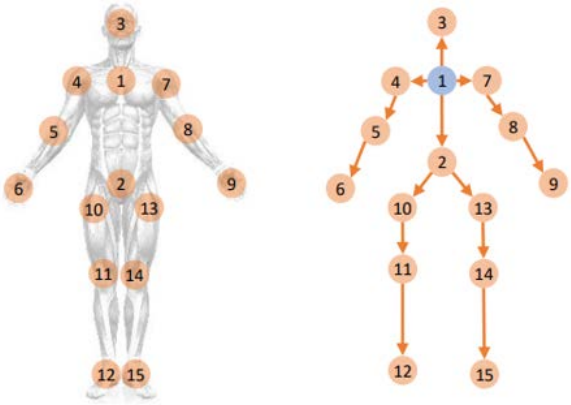


图 2 DGNN 有向图

论文：https://openaccess.thecvf.com/content_CVPR_2019/papers/Shi_Skeleton-Based_Action_Recognition_With_Directed_Graph_Neural_Networks_CVPR_2019_paper.pdf

代码：https://github.com/kenziyuliu/DGNN-PyTorch

3、AS-GCN

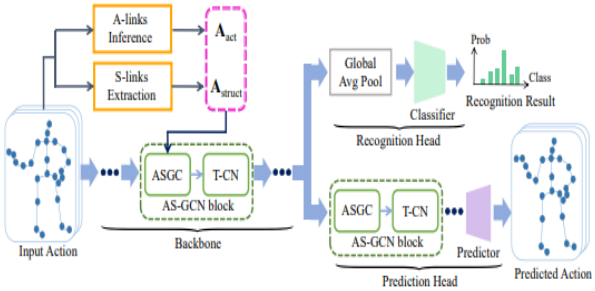


图 3 AS-GCN 结构图

工作：该论文提出了人体骨架的天然连接不适用于建模的观点，其工作侧重于将人体骨架图进行扩展，其采用了两种扩展方式。第一，对于人体骨架的天然连接，将其扩展到二次、三次等连接，也就是将本来在图中需要多次跳转的节点直接相连，扩展了图的结构，如图 3 所示。另一种方式是采用一个自编码网络，预先对其进行训练，得到一个由数据自学习得到的骨架图。该论文使用这两种扩展后骨架图取得了较好的结果。

论文：https://arxiv.org/pdf/1904.12659v1.pdf

代码：https://github.com/limaosen0/AS-GCN

4、2s-AGCN

工作：该论文主要针对 ST-GCN 的注意力机制灵活性不够，不能创造新的连接，即对于人体骨架图的建模仅依赖于定义的骨骼天然连接，如图 4 所示。由此，该论文提出了自适应图卷积，来自动的扩展骨架图的连接。另一个贡献在于，其不仅使用关节点的信息，进一步使用双流来利用骨骼的信息，将骨骼的长度和方向作为特征，来对人体动作进行建模，其方法补足了 ST-GCN 的两个缺点。

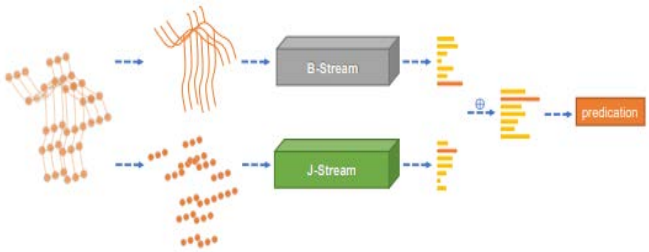


图 4 2s-AGCN 结构图

论文：https://arxiv.org/pdf/1805.07694v3.pdf

代码：https://github.com/benedekrozemberczki/pytorch_geometric_temporal

5、Shift-GCN

工作：该论文提出了一种将移位卷积应用到基于骨骼的行为识别中的方法，如图 5 所示，其利用1 × 1卷积算子结合空间 shift 操作，使得1 × 1卷积同时可融合空间域和通道域的信息，采用 Shift 卷积可以大幅度地减少参数量和计算量。该论文参考 Shift 卷积重新在骨架数据上定义 Shift 图卷积操作，通过 Shift 图卷积，该论文计算量相比于其他方法减少了近 10 倍。

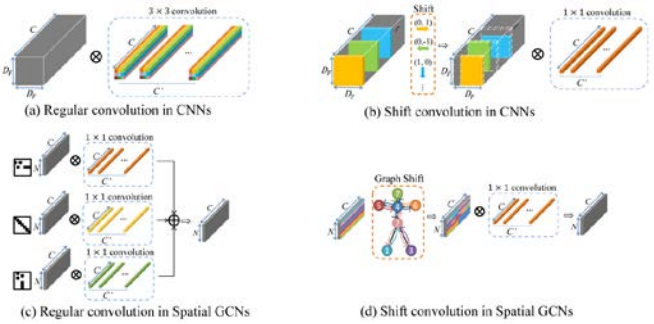


图 5 Shift-GCN 示意图

论文： https://openaccess.thecvf.com/content_CVPR_2020/papers/Cheng_Skeleton-Based_Action_Recognition_With_Shift_Graph_Convolutional_Network_CVPR_2020_paper.pdf

代码： https://github.com/kchengiva/Shift-GCN

6、ST-TR

工作： 该论文创新性的将 Transformer 引入到人体骨架动作识别领域，其使用 Transformer 来对人体骨架序列进行建模，如图 6 所示。本论文没有完全的将图神经网络舍弃，只是部分替换成 Transformer 模块，利用 Transformer 学习潜在的结构信息。该模型先经过三层 ST-GCN，然后将提取到的特征分别送入两个分支。第一个分支以 ST-GCN 为基础，使用 Transformer 替换掉其中的 GCN 部分。第二个分支同样以 ST-GCN 为基础，但是以 Transformer 替换掉其中的 TCN 部分。最终将双流网络提取到的特征共同进行 Softmax 分类。

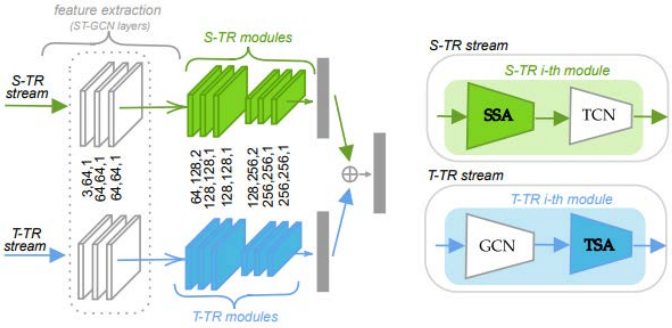


图 6 ST-TR 结构

论文： https://arxiv.org/pdf/2008.07404v4.pdf

代码： https://github.com/Chiaraplizz/ST-TR

责任编委 李策 贾同

张奇鹏

北京航空航天大学自动化科学与电气工程学院硕士研究生，研究方向为人体动作识别、计算机视觉。

王 田

北京航空航天大学人工智能研究院副教授，主要研究方向数据表示与挖掘，视频图像中运动信息提取，人机协同等。

多模态数据集

东北大学 贾同 武云鹤

多模态融合是指模型在完成分析和识别任务时处理不同类型数据的过程。多模态数据的融合可以为模型决策提供更多的信息，从而提高了决策总体结果的准确率，目的是建立能够处理和关联来自多种模态信息的模型，并已逐步成为研究热点。目前，多模态融合已经被广泛研究与应用于人的姿态识别、动作识别、显著性检测、行人检测及语义地图构建等领域。

随着多媒体技术的发展，图像数据急剧增长，如何在海量数据中高效提取出人们感兴趣、有价值的数据是目前亟待解决的问题。由于人眼在观察眼前图像时，只对图像中的一小部分区域感兴趣，因此，图像显著性检测领域具有十分重要的意义，其是指找到图片中人眼最感兴趣的区域。图像显著性检测研究对于图像分割、目标检测与识别、图像检索和图像压缩等多个领域都有重要的意义。目前，基于可见光的单模态图像显著性检测算法在特定测试数据集和简单场景下已经提升到较高水平，但是算法缺乏泛化性，颜色相近、跨边界，低光照度，图像包含噪声以及雨雪天气等复杂场景下显著性检测效果有待提升。近年来，随着多元传感器的发展，一些学者尝试融合多模态信息进行图像显著性检测。例如：获取具有丰富空间信息的深度图片与可见光图片共同获取显著信息或是利用对光照条件不敏感的红外图像与可见光图像结合起来提取显著区域。因此，本文重点介绍六种常用，基础的多模态显著性检测数据集，分别是 VT821 数据集，VT1000 数据集，STERE 数据集，GIT 数据集，DES 数据集，LFSD 数据集。

1、VT821 数据集

介绍：VT821 数据集是一个通用的多模态显著性数据集，每张图片的尺寸为480×640，常用于多模态显著性检测算法评测。VT821 是在不同场景和环境条件下收集的。其中室内场景包括办公室、公寓、图书馆等，室外场景包含道路、草坪、走廊、街道、建筑物等。正是由于 VT821 数据集中包含丰富的场景，使得该数据集更具有权威性，也使其成为多模态显著性检测领域常用的数据集之一。

VT821 数据集通过热成像仪(MAG32)和 CCD 相机(索尼 TD-2073)获取 821 对可见光-红外(RGB-T)高质量图像。与此同时，通过人工标注的方法获取对应的 821 张真值图。为了更好地促进不同算法的挑战敏感性能，VT821 数据集设立了 11 项显著性检测中常见的挑战，它们分别是：大的显著对象(BSO)、小显著对象(SSO)、多显著对象(MSO)、低照明(LI)、恶劣天气(BW)、中心偏移(CB)、跨图像边界(CIB)，前景与背景外观相似(SA)、热交叉(TC)、图像混乱(IC)和失焦(OF)。图 1 展示了 VT821 数据集中部分图片示例。

图 1 VT821 数据集图片示例。(a)可见光图像;(b)红外图像;(c)真值图

数据集地址

https://pan.baidu.com/share/init?surl=ksuUr3cr6_-fZAsSUp0n0w 提取码：9yqv

2、VT1000 数据集

介绍：VT1000 数据集含有 1000 个空间对齐的可见光-红外图像对及其对应的真值图像，其常用于评测多模态显著性检测算法的检测效果。VT1000 数据集收集了不同光照条件下的 10 种不同类型场景获取了 400 多种常见对象，比其他数据集得场景更加丰富。其中，10 种不同类型场景由室内场景和室外场景构成；室内场景包括：办公室、公寓、超市、餐厅、图书馆等。而室外场所包括公园、校园、街道、，建筑物、湖泊等。

VT1000 数据集通过硬件设备 FLIR SC620 来获取可见光-红外图像对。其中 FLIR SC620 设备中包含红外摄像机和 CCD 摄像机，这两个摄像机除了焦点外具有相同的成像参数，并且它们的光轴平行对齐，如图 2 所示。

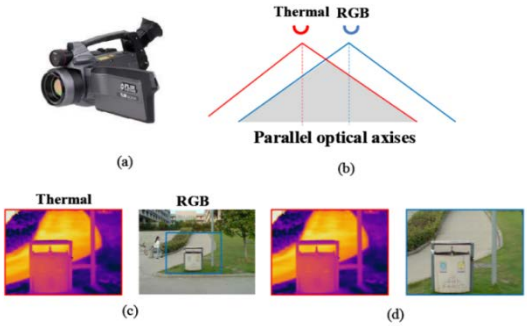


图 2 手动标记属性

由于硬件设备中的可见光相机和红外相机的成像参数基本相同，且光轴平行，因此可以通过静态或移动相机很好地捕捉其图像。两个模态图像之间的变换只是平移和缩放，这也使得具有不同模态的图像可以高度对齐，并且在边界中没有任何噪声，产生质量更高的可见光-红外图像对。

值得一提的是，VT1000 是在夏季收集的，这使得数据集中大多数对象的表面温度较高，许多图像出现热交叉，这无疑增加了对算法性能的挑战。与此同时，为了更好地评价不同算法的挑战敏感性，VT1000 数据集中还标注了 10 种挑战，分别是：大的显著物体对象(BSO)、小的显著物体对象(SSO)、多个显著物体对象(MSO)、低照度(LI)、中心偏移(CB)、交叉图像边界(CIB)，前景与背景具有类似外观(SA)、热交叉(TC)、图像杂波(IC)和失焦(OF)。图 3 展示了 VT1000 数据集中部分有挑战性的场景以及对应的真值图。

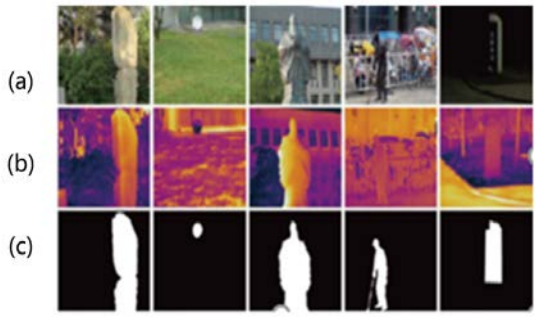


图 3 VT1000 数据集示例。(a)可见光图像;(b)红外图像;(c)真值图

数据集地址

https://pan.baidu.com/s/1i7gfrHoaaRuateMXBxvmMw 提取码：tb6l

图 4 STERE 数据集示例，从左到右依次为可见光图、深度图和显著性物体的真值图

3、STERE 数据集

介绍：作者利用 Flickr, NVIDIA 3D Vision Live 和 Stereoscopic Image Gallery 收集 1250 幅立体图像。每幅图像中最显著的物体均由 3 名用户进行标注，然后根据重叠的显著性区域对所有带标注的图像进行排序，选择前 100 幅图像来构建最终的数据集。这是首个可见光-深度(RGB-D)图像数据集。图 4 展示 STERE 数据集

中的可见光图片深度图片和显著性物体的真值图。

4、GIT 数据集

GIT 数据集由 80 幅彩色图像和深度图组成，这些图像是在真实家庭场景中利用移动机器人采集得到的。此外，每幅图像都是基于物体的像素级分割标注。图 5 展示了 GIT 数据集中的图片样例。

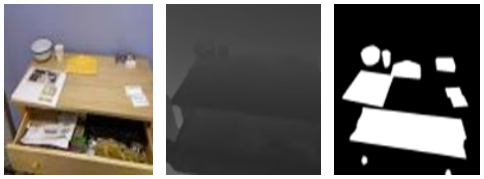


图 5 GIT 数据集中的图片样例，从左至右依次为可见光图、深度图和显著性物体的真值图

5、DES 数据集

介绍：DES 数据集由 135 张室内 RGB-D 图像组成，通过 Kinect 拍摄的分辨率为640×640。在标记该数据集时要求三个用户在每张图像中标记出显著性物体，然后标记图像的重叠区域作为显著性物体的真值。图 6 展示了 DES 数据集中图像样例。

图 6 DES 数据集中的图片样例，从左至右依次为可见光图、深度图和显著性物体的真值图

6、LFSD 数据集

LFSD 数据集是由 Lytro 光场相机收集的 100 张光场图像组成，包含 60 张室内场景和 40 张室外场景。为了标记该数据集，要求三个用户手动地对显著性物体进行分割，当三个结果的重叠率超过 90%时，将分割结果作为显著性目标的真值。图 7 展示了 LFSD 数据集中的图片样例。

图 7 LFSD 数据集中的图片样例，从左至右依次为可见光图、深度图和显著性物体的真值图

责任编委 沈沛意 李策

贾 同

东北大学信息科学与工程学院教授、博士生导师，智能感知与机器人研究所所长。研究方向为计算机视觉、模式识别与机器学习等。电子邮箱:jiatong@ise.neu.edu.cn

武云鹤

博士研究生，东北大学信息科学与工程学院，研究方向为计算机视觉。
电子邮箱： wuyunhe05@163.com

好 文 推 荐

长安大学团队“多目标跟踪”最新成果发表在 IEEE TPAMI-2021。

论文：ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, Mubarak Shah, Deep Affinity Network for Multiple Object Tracking, IEEE TPAMI, vol. 43, no. 1, pp. 104-119, Jan. 2021

多目标跟踪(MOT)在解决视频分析和计算机视觉中的许多基本问题方面发挥着重要作用。大多数 MOT 方法采用两个步骤，即目标检测和数据关联。目标检测是在视频的每一帧中检测感兴趣的目标，数据关联是在不同帧中对检出目标之间建立对应关系以获得它们的轨迹。近些年来，深度学习被越来越多地应用在目标检测领域内，并取得了巨大的进步。然而，用于跟踪的数据关联为计算不同帧中检出目标的亲和度仍然局限于单一的人工特征，如外观、运动、空间接近性等。

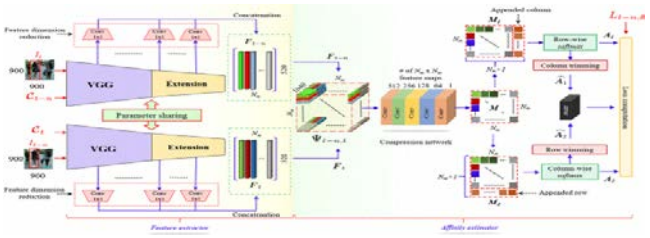


图 1 基于深度亲和力网络的多目标跟踪算法流程图

为此，长安大学团队利用了深度学习的表示能力。提出了一个深度亲和网络(DAN)，如图 1 所示。它以端到端的方式在一对视频帧中联合学习目标物体的外观特征，并直接估算检出目标间的亲和力。该网络可学习到考虑了物体及其周围环境在多个抽象层次上的分层特征，基于这些特征，DAN 网络可估计出不同帧下检出目标之间的亲和度。此外，我们为 DAN 设计了一个前向-后向损失函数，使其具有表征两个帧之间目标出现或消失的情况。DAN 用一个具有共享参数的主干网络对物体外观进行建模，并利用浅层和深层特征估计物体的亲缘关系。

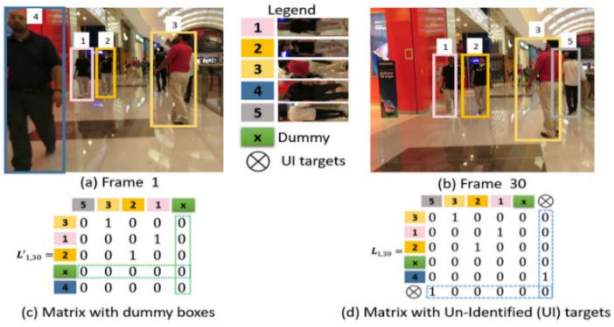


图 2 帧 1 和帧 30 之间的数据关联矩阵示意图

如图 2 所示为构造数据关联矩阵的方法，在图中，帧 1 和帧 30 各包含 4 个被检测到的人，相当于 5 个不同的身份。其中图 2(c)为用于具有一行和一列虚拟包围框中间矩阵的构建，在图 2(d)中，关联矩阵增加了额外的行和列，标记为未识别的目标。增强的列表示当前跟踪的对象离开视频，增强的行表示进入视频的新对象。使用这个规则，深度亲和网络能够解释多个对象离开和进入视频。本文最终使用图 2(d)所示的真实关联矩阵的形式来训练深度亲和网络。

本文利用所提出模型的高效亲和力计算将给定视频帧中的对象与多个先前帧中的对象相关联，以使用 Hungarian 算法生成可靠的轨迹，以实现准确的在线多对象跟踪，如图 2 所示。所提出的方法在流行的 MOT15, MOT17 and UADETRAC 挑战数据集上以每秒 6.3 帧的速度实现目标跟踪，同时在大多数评估指标上超过了现有的领先方法。本文在在线多行人跟踪方面具有明显优势。

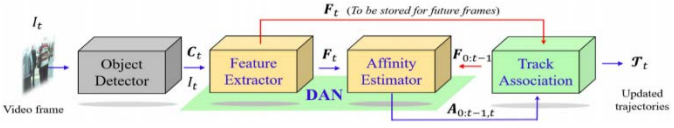


图 3 深度关联网络的部署流程图

责任编委 樊鑫 贾同

好 文 推 荐

西安电子科技大学团队“三维点云目标检测”最新成果发表在 IEEE TIP-2021。

论文：Mingtao Feng, Syed Zulqarnain Gilani, Yaonan Wang, Liang Zhang, Ajmal Mian, Relation graph network for 3D object detection in point clouds, IEEE TIP, vol. 30, pp. 92-107, Oct. 2021

随着三维传感器，如三维扫描仪、深度相机和光场相机等的问世，三维点云已经在许多领域得到越来越多的应用，如机器人、自动驾驶、城市规划，古建筑维护等领域。精确检测点云中的三维物体是移动机器人自动规避障碍物、规划路径和与物体交互的核心问题。将点云转换为标准形式，如深度图像、多视图或体素，已经成为使用卷积神经处理 3D 数据的流行方法。然而，将深度网络直接应用于原始点云的坐标进行三维物体检测还没有得到广泛的研究。由于三维点云的不规则性和稀疏性，三维目标检测的进展远远滞后于二维目标检测。

为此，西安电子科技大学团队提出基于关系图网络的三维目标检测算法，如图 1 所示。本文首先扩展了二维目标检测方案，提出了一种精确的几何中心估计和三维候选框回归算法。但是，与物体中心像素被其他像素包围的二维图像不同，三维物体的几何中心通常在远离物体表面点的空间中。此外，在复杂场景中的三维物体通常是局部扫描的，且具有一定的噪声。因此，基于点的语义特征，测量物体表面点到其几何中心的距离的偏移值很难直接回归。几何中心预测不准确会导致下游三维候选框生成器产生误差。因此，从物体表面点中学习更多的鉴别特征，更准确地计算中心位置，是正确回归三维候选框的关键。本文首先引入了一种联合预测伪几何中心(或简称伪中心)和方向向量的策略，从而为 3D 候选框回归提供了一个双赢的解决方案。我们预测接近三维物体真实几何中心的伪中心，并为每个表面点分配指向该几何中心的方向向量。这些向量的大小和方向相互协作，进一步提高了三维边界框候选的准确性。

由于在复杂场景中的物体是随机放置并紧密连接

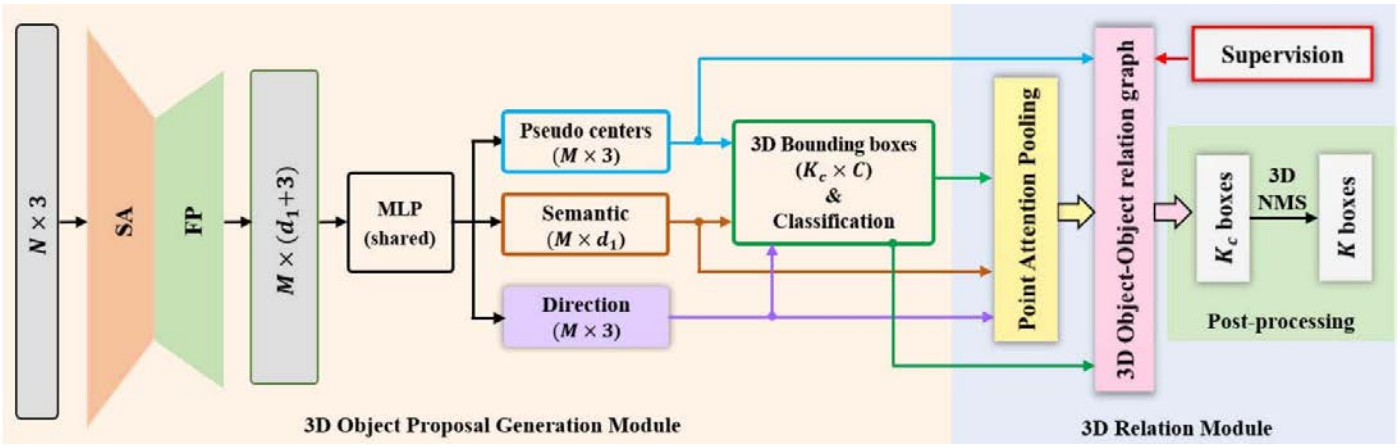


图 1 基于关系图网络的三维目标检测算法流程图

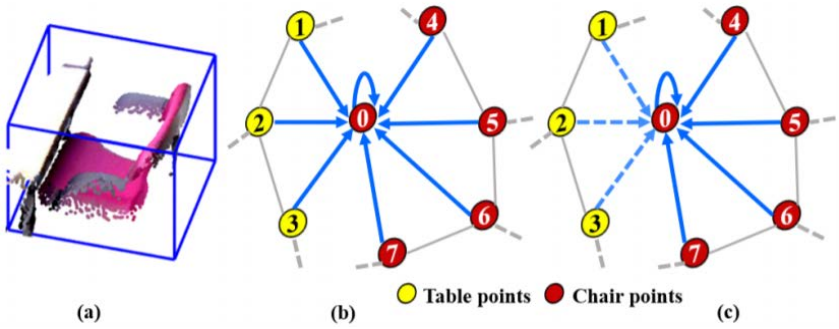


图 2 候选框内点注意力机制示意图

的(例如，椅子的座位部分可能在桌子下面)，一个 3D 候选框通常包含来自不同物体的部分。本文引入了点注意池化方法来提取每个三维候选框的统一外观特征，并将其与位置特征一起用于定义关系图的节点，如图 2 所示。本文提出的点注意力池化方法利用从三维候选框中获得的信息，同时建模内部点的语义、空间和方向关系。这使得属于同一对象点的拉力和不同对象间点的推力共同作用。

回归 3D 边界框通常会导致候选框的冗余。去除重复数据的一种简单方法是使用具有阈值的 3D 非最大抑制(NMS)。与 3D NMS 相比，去除重复项更直观的思路是利用复杂场景中不同对象之间的关系。例如，椅子经常靠近桌子，电脑经常放在桌子上。然而，由于在复杂三维场景中，属于不同类别的物体之间的距离是任意的，而且物体的数量和大小也不同，因此，三维物体之间的对象关系很难建模。本文提出了一种有效的三维关系模块，如图 3 所示。该模块构建了三维候选框之间的三维目标-目标关系图，用于特征增强，实现准确的目标检测。将上述关系图插入到主框架中，并通过最小化任务特定损失(如 3D 候选框回归损失、交叉熵损失、方向特征损失)的方式进行无监督学习。

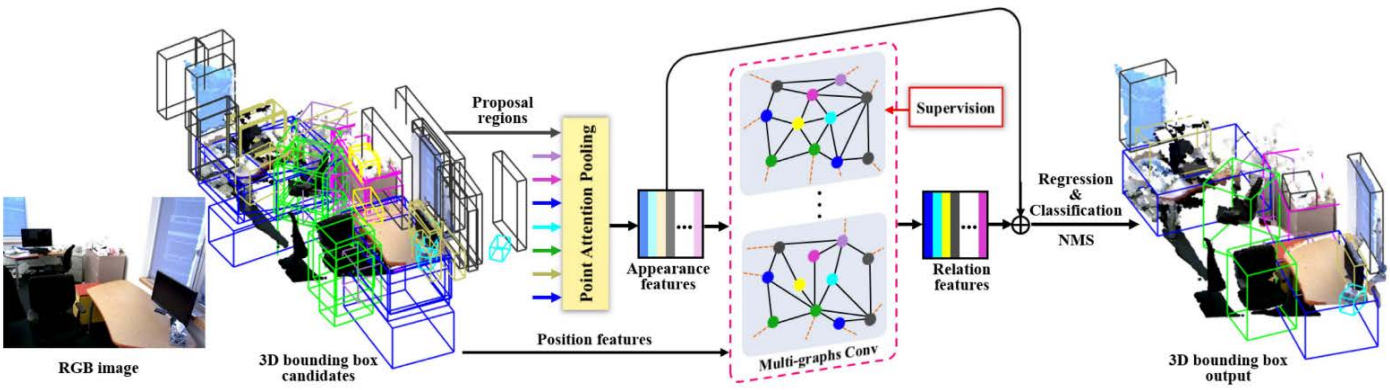


图 3 基于目标-目标关系图网络的算法流程图

责任编委 樊鑫 贾同

好 文 推 荐

台北科技大学的“DSNet: Joint Semantic Learning for Object Detection in Inclement Weather Conditions”最新成果发表在 IEEE TPAMI 2021。

论文：Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. DSNet: Joint Semantic Learning for Object Detection in Inclement Weather Conditions, IEEE TPAMI, 43(8): 2623-2633, 2021

目标检测在自动驾驶汽车中起着至关重要的作用，因为它不仅可以确定每个目标所属的类别，并在给定的图像中定位目标，还可以帮助系统在复杂的驾驶环境中实现安全导航。最近几年，许多目标检测方法和技术被引入，并取得了令人印象深刻的性能。现阶段，尽管这些检测方法在正常图像中表现出良好的性能，但它们通常无法在恶劣天气条件下检测到物体，尤其是在雾中。而雾是驾驶场景中最常见的天气现象之一，因此如何提高检测方法在雾中天气的检测效果成为研究的热点。

针对这一问题，本文提出了一种名为 DSNet 的目标检测网络。如图(1)所示，DSNet 采用了目标检测方法 RetinaNet 作为骨干网(表示为检测子网)，并在此基础架构上提出了特征恢复模块，构建了用于增强能见度的恢复子网。下面进行详细描述。

(1)检测子网。DSNet 采用 RetinaNet 作为检测子网的主干方法。RetinaNet 提出的损失函数 Focal Loss 可以减少易于分类样本的损失，并在训练中更加注重学习难以分类的示例，从而实现类与类之间的平衡。此外，RetinaNet 采用的特征金字塔网络，提供自顶向下和横向连接的路径，可利用丰富的语义层构建更高分辨率的层，从而显著提高在雾霾存在时小物体检测的准确性。

(2)恢复子网。恢复子网负责生成与检测子网共享的共享特征，同时恢复带雾图像的特征，以提高能见度较差条件下目标检测的准确性。该子网络在改造后的大气散射模型的基础上完成目标，使用了两个模块：1)共享模块；2)特征恢复模块。

大量实验表明了所提方法在去雾方面的优越性。

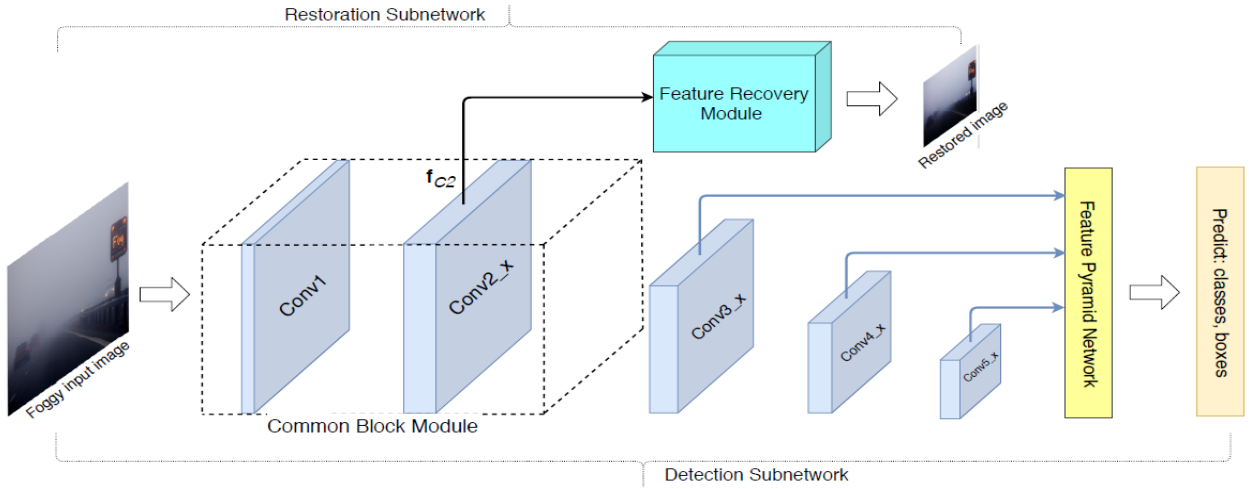


图 1 DSNet 方法流程图

责任编委 贾同 樊鑫

征文通知

1 会议征文

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

2 期刊征文

计算机视觉领域近期相关期刊专刊的征文通知如表 2 所示，包括 IEEE Journal of Biomedical and Health Informatics, ACM Transactions on Multimedia Computing, Communications and Application 和 Image and Vision Computing。

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。

第四届 PRCV 将于 2021 年 10 月 29 日至 11 月 1 日在北京国际会议中心举行，由北京科技大学、北京交通大学和北京邮电大学共同承办，中山大学、清华大学协办。本届会议将主要汇聚国内从事 PRCV 理论与应用研究的广大科研工作者及工业界同仁，共同分享我国 PRCV 领域的最新理论和技术成果，为大家提供精彩的学术盛宴。现向广大科技工作者公开征集高质量、原创性的优秀论文。会议论文集将由 Springer 出版社 LNCS 系列出版，并被 EI 和 CPCI-S 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
ICASSP 2022	2022.05.22-27	Singapore	2021.10.01	http://iab-rubric.org/fg2021/
ICLR 2022	2022.04.25-29	Online	2021.10.06	https://iclr.cc/
AISTATS 2022	2022.03.30-4.1	Valencia, Spain	2021.10.15	https://www.aistats.org/aistats2022/
CVPR 2022	2022.06.21-24	Louisiana, USA	2021.11.17	http://cvpr2022.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
JBHI	Skin Image Analysis in the Age of Deep Learning	https://www.embs.org/jbhi/special-issues-page/special-issue-on-skin-image-analysis-in-the-age-of-deep-learning/	2021.11.01
TOMM	Affective Services based on Representation Learning	https://dl.acm.org/journal/tomm/special-issues	2021.11.15
IVC	Deep Learning Techniques Applied to Faces	https://www.journals.elsevier.com/image-and-vision-computing/call-for-papers/deep-learning-techniques-applied-to-faces	2021.11.30
IVC	Video Computation and Reconstruction in Digital Twins	https://www.journals.elsevier.com/image-and-vision-computing/call-for-papers/-video-computation-and-reconstruction-in-digital-twins	2021.12.20

心底无私视界宽∞徐光祐教授专访

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者。

《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。

2021 年 4 月，《CCF-CV 专委简报》委托贾熹滨教授通过微信交流方式专访了清华大学徐光祐教授。为能保持本次访谈的原汁原味，我们采用问答的形式，并整理了徐教授的专访记录。

问：1、我们了解到，您是 1963 年毕业于北京清华大学自动控制系，毕业后直接留校任教。作为人机交互和多媒体计算领域的知名学者，在人工智能、计算机视觉领域都很有建树。您的成果包括早期底层 “软磁盘驱动器”，在信息处理层和高层理解方面的“基于知识的多传感器信息处理和理解”、“多种视觉传感器系统综合理解技术”等，以及平台研制“TH - DMDS 数字多媒体开发系统”、“多媒体群件的研制与应用”等一系列国家级省部级奖励。您能否和我们分享回顾一下您的研究经历以及这些重要成果背后的故事?

徐教授：以下是我对上述：“视界专访”部分问题的思考和回答。

■ 我为什么选择计算机视觉作为研究方向

清华大学计算机系的前身是清华大学自动控制系，它是原高教部为了给国家的两弹一星计划培养人才而建立的新系。我系的大批毕业生就分配到与两弹一星相关的研究所工作。我所在的教研室原来属于“飞行器控制”。后来由于技术发展的需要，新增设了“自动控制元件”教研室。文革以后，相关方向的人才改由相关工业部门所属的高等学校培养，学校的院系也进行了相应改革。我系在 1977-1984 年曾经改为电子工程系，又于 1984 年改名为计算机科学与技术系。在电子工程系的时期，学校院校的教研室是以产品为方向。我所在研究室是“计算机外部设备”。这就是我从事软磁盘机研究的原因。后来我们发现以工业产品为教研室的方向存在一个重大的问题：教研室的方向事关人才培养的方向，它不但需要稳定，而且应该是有长远发展前途的方向。但是，以产品为导向就难以做到这一点。以计算机的外存设备为例，它就经历了软盘存储器、硬盘存储器、光盘存储器等不同类型。而且，每种类型的关键技术又有很大差异，缺乏可继承性。因此，研究室应该以学科为方向，这样才可能有长期稳定的发展方向，这是取得高水平研究成果的前提。因为学术研究取得高水平成果的前提就是在一个方向上进行长期锲而不舍的钻研。理论上讲，越是基础的学科方向越可保持稳定。例如，数学物理。但是在现实环境下，特别是在发展迅速的计算机

科学与技术领域，在不同发展阶段需要各种不同专长，而且能迅速适应工作的人才。为此，在我系改为计算机科学与技术系以后，我们教研室就改名为“信息处理与应用”教研室。

“信息处理与应用”教研室主要从事语音和图像信息的处理和识别研究。这就是以后我们研究室从事多媒体信息处理研究的基础。应该说，“信息处理与应用”是一个相当笼统的名称，虽然具有了广泛的可适应性，但难以用来指导和推动实际的教学和科研工作。反观与计算机系同在信息领域的其他两个系“清华大学无线电系”和“清华大学自动化系”，早已分别在学部委员吴祐寿教授领导下建立了“图像信息处理和识别”教研室，因为他们在卫星遥感图像处理方面已有良好的基础。而自动化系则在学部委员常侗教授领导下成立了“模式识别”教研室。这样迫使我们必须在图像信息领域中另辟蹊径，寻找新的方向。作为图像领域的后来者，我们发展的机缘来自两方面。一方面是我在 1982-84 年幸运地被公派到普渡大学的美国科学院院士傅京荪 (K.S.Fun) 教授团队作访问学者。傅京荪教授被尊称为句法模式识别之父，在他去世后国际模式识别学会特设了 K.S.Fu 奖（Prize），来纪念他的不朽贡献。在普渡大学的访问期间，我聆听了傅教授的相关课程，开始从事图像处理和识别方面的工作。在回国前，傅教授出于对中国科技事业的支持，希望我和他的另一位访问学者 (蔡自兴教授) 一起撰写《人工智能及其应用》一书，以便在清华出版社出版。这本书曾被国内不少高校采用为人工智能方面的教材，推动了国内人工智能领域的学习和研究。傅教授是在 1985 年代表普渡大学申请美国当时的 6 个重大科技发展中心(每个项目 1500 万美元)并得到成功时，不幸突发心梗而去世的。这个项目的名称就是“计算机支持的集成制造系统”。这项技术已经在当今的先进制造系统的发展中起到重要的作用。傅教授虽然已经为模式识别领域做出了杰出的贡献，但在科学探索的道路上从来没有停息，直到生命的最后。这是我所见到的其他美国学者无法比拟的。

问：2、我们想请您分享一下您的求学和学术研究经历，讲讲您曾经经历的故事，遇到过的困难，以及有趣的轶事。

徐教授：在这里我要感谢傅京荪教授把我引入了人工智能和模式识别领域。此外我想借此机会与大家分享一下与此相关的小故事：

- 如上所说傅京荪教授对推动我的事业发展有很大的影响，但我到他那里的访问研究却有很大的偶然性。因为那时我对国际上在模式识别领域的发展状况还缺乏基本了解，虽然得到了去美国做访问研究的机会，但对到哪个大学、哪位教授那里去访问，我心中一片茫然。无奈之际，只好去图书馆查阅相关的国际学术刊物。在这些刊物中发现，K.S.Fu 的出现频率很高。我就猜想他一定很有名，同时从名字看应该是一位华裔。所以，我就贸然向他发了一封申请书，说我对他的研究很感兴趣，也读了不少他的论文。所以，想去他那里做访问研究。说实在的，其实我对傅教授的研究内容还很不了解。与此同时，我还向美国田纳西大学的冈萨雷斯（Gonzalez）教授发了类似的申请信。因为，那时中国刚出版了一本他所编写的数字图像处理教材，使他在中国很出名。虽然，这两位教授都接受了我的申请，但我最后选择了傅教授。

- 那么我去美国访问之际，国内在图像处理和识别的基础如何？记得当时（1982 年）我们教研室的全体老师都开始在教研室公用的美国惠普公司生产的 PDP-11-23 小型机上学习计算机编程，从 Pascal 语言开始。当时，在计算机系，不但一般的教师不会编程，就是计算机专业、从事计算机研制的老师也不用计算机和编程。那时每位教师分配到的内存仅仅是 2 MB（兆字节），在现在看来这简直是玩笑。2 MB 能干什么？我刚到普渡大学时，分配给我的内存是 20 MB，申请后可增加到 40 MB。当时，我还感到已经有很大的改善了。

其实，这也难怪，因为那时一般所处理的图像分辨率是128x128，如能到256x256，那就算是高质量了。那时，根本没有图形显示器，都是字符显示器。图像的灰度等级，主要是靠字符打印机上对不同字符组合来分辨不同的灰度。在普渡大学傅京荪教授与另一位A.V.Kak 教授（他与著名的 Prof. Azriel Rosenfeld 教授合著了《Digital Picture Processing》一书）合作的实验室中有一台使用热敏纸的打印机才能打印灰度图像，主要供正式发表论文时用。

问： 3、在您的科研学术经历里是否遇到过根据国家发展需要调整科研方向的问题？还是您会主动地去发现热点，去调整研究方向呢？您认为一名科研工作者以什么样的精神或态度从事科研，才能跟上时代的发展需要？

徐教授：以下是我对上述：“视界专访”部分问题的思考和回答。

■ 计算机视觉是基础性的学科方向

我 1984 年回国以后，就开始准备和讲授计算机视觉。之所以选择计算机视觉作为我们研究室的一个学科方向，这既是为了适应人工智能的发展需要，也因为计算机视觉是比图像处理更为深刻和根本的学科。从根本上看，图像处理和识别只代表对传感器的数据的处理和识别，而视觉本身就具有思维的特征。从视网膜输入的图像到人们大脑的视觉感知是一个深刻的认知过程。有的学者甚至提出视觉思维的观点。计算机视觉是一门富有挑战性的学科。人们已经认识到人类信息的 70-80% 是来自视觉，也更说明了它的重要性。

但是学科要健康地发展离不开应用的推动。应用从来就是学科发展最根本的动力。因此，一定要发现当前的应用需求。智能机器人和多媒体技术是当时计算机视觉的最迫切的应用领域。

- 我们研究室选择计算机视觉作为重要的学科方向以后，还需要寻找是否有国家需要的应用项目。幸运的是那时正值国家的第 6 个五年计划期间，国家正在大力发展工业和智能机器人。而国防领域同样需要智能机器人，特别是移动机器人。于是，我们与系内的“自动控制教研室“（后来的人工智能教研室）合作参与了国家的各种机器人重点攻关项目，其中包括国防科技攻关项目——军用移动机器人的研究。我们负责其中的视觉导航和侦察任务。这就为计算机视觉找到了明确的应用领域。我国的国防移动机器人项目是以美国的 DARPA 计划的 ALV (Autonomous Land Vehicle)，即自主移动车项目为参照。它与目前市场上正在开发的自动驾驶汽车（automatic vehicle）有本质的区别，因为 ALV 在战场上可能面临的是复杂而且多变的环境。也就是说 ALV 要面对的是未知的、非结构化的动态环境（即动态的上下境—dynamic context），而目前市场上的自动汽车的上下境是可以通过卫星定位和已知的，环境模型在事先已知或固定的。

- 朱志刚的博士论文就是以 ALV 为应用背景的一个富有挑战性的前沿科研课题，这也为他获得全国优秀博士论文打下了基础。从朱志刚的博士论文的研究可看到，如要取得创新性的成果就需要面向应用，挑战前沿的科学难题。如果不结合实际应用，就难以明确研究的方向，从而取得好的科研成果。

■ 计算机视觉与多媒体研究的相互促进和发展

如上所述，在我们教研室以“信息处理与应用”为学科方向的阶段，国际上由于在半导体集成电路的迅速发展，使计算机的计算能力和传输能力极大地提高，使得计算机有可能实时处理语音和图像甚至视频数据的能力，从而促进和推动了多媒体技术兴起。我们在国内较早地认识到这个方向的重要性，率先开展了多媒体的研究，并努力推动国内在此领域的研究和发展工作。其中包括由我们系主动向国家计委提出了在第 6 个国家五年计划中列入多媒体的攻关项目的建议。并且通过中国计算机学会和中国图象图形学学会，1992 年在北京发起和主持了多媒体领域的国内学术会议，即全国多媒体技术学术会议（NCMT，National Conference on Multimedia Technology)。从 1992 年开始这个会议每年召开一次。这个会议有力地推动和促进了我国多媒

体领域的发展和研究。

多媒体显然是一项对计算机行业的发展和应用具有重要意义的技术。作为工科院校，无疑应该大力发展和研究。但如何与此项结合，开展高水平的学术研究是一个重要和根本性的问题。初期为了建立基本的研究环境，我们参加和从事不少语音、图像与视频信息的计算机接口电路的研究。其中主要的工作是相应的接口卡。但这项工作虽然重要，在很大程度上受制于半导体芯片，特别是模-数转换芯片的性能以及计算速度。这是我们所无法左右的事情。为了明确研究方向，我们对发展多媒体技术的关键问题进行了分析和梳理，其中主要包括以下 4 方面：多媒体数据的压缩，存储，传输（即分布式多媒体技术），多媒体文件的编著（即多媒体文件格式的编辑和著作）。

当时最迫切的是多媒体文件的编著软件，即编著和开发工具。因为当时只有文本结构的文件形式，而多媒体文件需要超文本结构。缺乏多媒体著作工具就难以开发多媒体文件和应用，从而阻碍了多媒体技术的应用开发。为此我们首先开发了一个多媒体著作系统，以促进多媒体应用的发展。但这是一项与应用环境高度相关的工作，需要进行大量的开发和服务性质的工作。因此，虽然我们在国内领先开发了一个成功的多媒体著作系统，并得到了重要的应用。由于这项工作虽然重要但不适合在学校研究，我们只能放弃，于是改为集中精力研究分布式多媒体技术。

我们发现分布式多媒体技术具有重大的应用前景。特别是计算机主持的协调工作（Computer Supported Collaborative Work, CSCW）将帮助人们克服在时间（时区）和地理位置上的差别，进行协同工作，从而可极大地提高人们的工作效率和质量。它在一些特殊的、需要各种不同专业人员相互紧密配合工作的行业，例如飞机设计行业中更是具有不可取代的作用。开始时，我们也开展和参加了分布式飞机设计系统的研发，参加了相应的高技术研发 863 项目。但后来发现，我国的飞机设计软件本身是进口的。我们无法接触到这样的软件的核心，也就无法进行分布式系统的研究。为此，我们就结合当时国家急需的远程教育，着重开展了 CSCW 在远程教育中的应用，即远程教育中的智能教室的研究。这方面的研究实际上为我们后来进行普适计算、人机交互领域的学术研究打开了宽广的大门。

因为，从学术研究的角度来看，CSCW 的本质就是借助于分布式计算机系统，通过人机交互实现不同时区和地点人们之间的合作。后来移动通信的发展，又使交互空间从仅限于计算机的面前扩展到人们的生活空间。所以，从本质上来说，分布式多媒体，CSCW，普适计算都是用于支持人机交互。而视觉和听觉是人类最重要和基本的感知系统。这就不难理解，为什么计算机视觉无论在分布式多媒体、普适计算、人机交互研究中都是一门不可取代的基础学科。关于上述技术与人机交互以及计算机视觉之间的关系的分析，可参考我撰写的“普适计算教育部重点实验室”的申请报告。

问：4、 目前学界的工作成果越来越丰富，也吸引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？

徐教授：进行高水平的学术研究，首先需要有明确和稳定的基础学科方向，并在此方向上进行长期锲而不舍的钻研。另一方面，为了能紧跟科技发展的前沿，还需要不断地更新和调整在此基础学科上的具体课题的方向。这样才能既保持在基础学科研究上的稳定性，又能始终站在科技发展的前沿和适应国家发展的应用需求，从而使学科方向的研究保持活力。

问： 5、您的研究从多媒体个人计算机到人机交互乃至普适计算模式下的人机交互，从动作识别理解到从人机交互中的体态语言理解至以人为中心计算机视觉，能通过您的科研经历，回顾一下中国人机交互、普适计算领域学术发展历史吗？还有哪些值得回忆的、关于该领域有趣的轶事？另外，您如何看待这一领域的发展：过去，

今天与未来?

徐教授：在这里我还想介绍一下，为什么虽然我们从开始 CSCW 领域的科研工作开始，就逐渐地认识到人机交互作为基础学科的重要性，而我们却迟迟不敢在教研室的名称上加上“人机交互”的字样？我想与大家分享我以下的经历：

首先，我对人机交互的了解是从 1992 年 12 月-1993 年 6 月在美国 Beckman Institute, University of Illinois at Urbana Champaign（即 UIUC）的黄旭涛（Prof. Thomas Huang）做访问教授时开始的。Beckman Institute 是美国伊利诺伊大学香槟分校（UIUC）所属的一个从事基础研究的研究所。它开展了从生理机制上研究人类视觉这样的基础研究，还包括了人机交互的研究。虽然具体所涉及的是人脸识别技术，包括参加在美国自然科学基金会发布的人脸识别测试数据库上的竞赛，但与此同时还专门成立了人机交互研究小组。那时总的感觉是人机交互领域涉及的面很广，其中很大的部分是心理学等方面。而在计算机和信息领域中，对人机交互的研究虽然已有较长的历史，比如在国际上历史最久和著名的国际学术会议 CHI (Computer Human Interaction)，是从上世纪 50 年代就开始有的。但出席会议的主要是来自心理学和人因工程学方面的专家。而人因工程学是一门与应用密切结合，而我们所不熟悉的学科。对此，我们无力介入。

为了开展人机交互方面的研究，我们也探索了与心理学方面专家合作的可能性。为此，我们与科学院心理研究所联系和探讨了进行合作研究的可能性。其中包括相互参加对方的研讨会，博士生学位论文的评审和答辩会等。总的感觉是：这是一个重要和需要合作的领域，但由于分别属于不同的领域（例如，计算机和信息处理领域总的来说是一个工科的范畴，而心理学是属于理科的范畴。对博士论文的评审标准就有根本的差异：工科博士论文首先关心的是创新性，所提出的方法不一定是唯一的，但应该是最好的；而理科要求的首先是正确性，即是否符合科学规律本身，它应该是唯一的）。所以，从我们工程的角度来看心理所的博士论文所追求或最后得到的结论方法大多看起来是很简单和显而易见的。虽然心理学得到的结论是符合规律和唯一的，但离实际应用有较大距离。所以，虽然双方都觉得从学术来说，应该相互开展学术交流，但从实用的角度来看，双方还需要有更多的相互渗透和了解，以便逐渐建立合作的基础。可喜的是，我们教研室的蔡莲红教授在语音的情感识别领域与心理所的专家进行了很成功的合作，取得了双方都满意的成果。

这样我们就面临一个两难的境地，既觉得应该早一些举起“人机交互”这面大旗，因为它可能代表了信息处理领域重要的学科方向，但又顾虑教研室被误解为是属于心理学领域的研究所，以至学生毕业后找不到工作。这就是为什么我们教研室的名称前面加上“媒体集成”。至于“媒体集成”的名称是通过与美国南加州大学（USC，University of South California）的 Neuman 教授在学术交流中了解到他们那里有一个美国自然科学基金会的“媒体集成”重点实验室受到了启发。于是我们教研室的名称成为“媒体集成与人机交互”。因为前面的“媒体集成”明确了这是我们开展人机交互的基础。而时至今日，人机交互在人工智能领域中的重要性已开始为人们所理解和接受。 所以，把人机交互纳入人工智能研究院也是顺理成章的事情。

问：6、 您知道深度学习现在非常热，特别在视觉研究领域，出现了一系列卓有成效的成果，三位深度学习大家也是同时获了图灵奖。但相比经典的计算机视觉研究方法，深度学习方法缺乏可解释性又使得深度学习难以突破现在的框架设计和调参，似乎达到了深度学习天花板。我们了解到，您非常关注中国计算机视觉未来的发展，关注在深度学习的冲击下模式识别和计算机视觉如何发展、深度学习是否能解决及触及计算机视觉的根本学术问题。能不能对这些问题，在此先和视界专访的读者们分享一下您的一些看法？在这方面您给现在的研究人员有什么建议吗?

徐教授：关于这部分问题的讨论：我想主要与各位探讨

一下“深度学习热”的情况下，如何开展高水平的计算机视觉研究，或者可能需要探讨一下深度学习对计算机视觉研究的影响。

在这里我不想对深度学习做一般性的学术评价，只是做了一个供参考的初步思考。

■ 首先可能需要简单地探讨一下“计算机视觉”与“模式识别”或“图像识别”之间的关系和区别。

“图像识别”与“计算机视觉”密切相关，或者可以说是“计算机视觉”的基础之一，但这二者还是存在重要甚至是本质的区别。首先回顾一下使计算机视觉成为一门独立的学科的标志性成果是什么？通常认为 Roberts Rob 在 1965 年发表的博士论文[Rob 65]是计算机视觉研究中的开拓性工作。他研究了根据二维的图像来理解由多面体积木块所构成的三维景物（以后常称为积木世界）的方法，也就是完成了从二维图像到三维景物的推理和理解，这就突破了图像处理和识别的局限性，在[威 07]中，作者进一步认为视觉具有思维的特征。虽然目前图像处理和识别中也可能包括深度图像，但目前为止还不涉及相关的推理和理解。

与此相关的一个问题是在深度学习方法高度发展和取得巨大成功的情况下，大学中应该如何进行计算机视觉课程的教学。为了使学生能快速适应毕业后的工作，需要加强深度学习方法的内容，这是学生所欢迎的，马上就可用的内容。但从为了学生能够对计算机视觉进一步学习和研究的角度来看，需要加强对计算机视觉中的基本概念和理论的介绍和探讨，但这可能是一项吃力不讨好的任务。如何在这两者之间取得平衡是一个重要的问题，与此相应的是迫切需要合适的教材作为支撑。

■ 深度学习方法对计算机视觉研究的促进

毫无疑问，如[Ser 19]中所指出的：“在 2012 年至 2015 年的短时间内，基于神经网络的解决方案已经克服了许多计算机视觉挑战。最显著的是 ImageNet 大规模视觉识别挑战（ILSVRC）。2012 年，最先进的解决方案基于传统的图像处理和模式识别技术，实现了 26.1% 的前 5 位错误率。同年，AlexNet 体系结构的开创性工作令人印象深刻地将错误率降低到了 15.3%。三年后，微软研究小组的研究人员用他们的残差网络实现了人类水平的绩效，这是一个令人惊讶的 5.71%的前五名误差。”[Sze 15][Zha 16]。

■ 深度学习方法有明显的局限性[Dro 21]

首先，目前的图像识别只是局限于模式识别的范畴，而计算机视觉需要从识别结果进一步推论识别结果所包含的语义，也就是需要理解。所谓的理解就离不开当时的具体状况，就是上下境（context）或目前通常称之为上下文（需要说明的是，这里的上下境与在语音处理和识别中的上下文的差别是它应该包括时间、空间两个维度，而不是只有“上下文”一个维度）。

图像识别如要进一步提高到理解的层次，就需要理解识别结果所包含的语义，这就离不开当时的情境，即上下境。现在计算机视觉界已开始普遍认识到上下境（context）在视觉理解中的重要性。视觉推理中不但需要引入上下境，而且需要能应用动态上下境的情况，也就需要具有觉察上下境的能力。这对于理解人体动作（体态语言）尤其是这样。

在这里我想引用一下中国科学院院士清华大学教授张钹写的《人机交互中的体态语言理解》[徐 14] 一书的书评中，关于对智能，语义与上下境之间关系的一段话：

“……，图灵提出图灵机的概念时使用“可计算实数”来定义计算机的可计算性。在论述机器（计算机）智能的概念，如图灵测试时，也只是从机器的表观行为上定义“智能”，与智能的本质并无关系。事实上，在“智能”的框架下，计算机依然只扮演数字计算的角色。正因如此，传统的计算理论总是避开“语义”这一难题，把理论建立在与“内容无关”（meaning independent）的假设之上。但当讨论“体态语言理解”时“语义”却成为一个绕不开的话题。计算机能否从用户行为中获得其隐含的语义，至少需要解决三个层面上

的问题：（1）识别语音、图像、表情、手势等信号，这是模式识别问题；（2）推断隐藏在信号背后的语义；(3) 估计受众对这些信息的反应，以及所产生的影响。就“理解”而言，需要解决的科学问题是，依据信息发送者发出的信号，预测和构造他本身以及信息接收者的认知模型。显然，如果没有周围物理与社会环境的信息，没有用户心理状态的知识，仅仅依靠建立在数学模型基础上的数据处理，是不可能完成这一任务的。本书介绍的上下境模型以及觉察上下境的计算范式，是一个很好的解决方案。它把从底向上的数据驱动与自顶向下的知识导引结合起来，即将传统的信息处理与人工智能的方法结合起来。”

■ 当前在计算机视觉研究需要中引入觉察上下境机制的应用领域在哪里?

学科的研究需要有巨大应用需求的推动才有生命力。这样的应用需求是什么？这是一个挑战性的课题。

欧盟在 2020 年前在信息领域的一个具有重大需求的研究项目是 Ambient Intelligence，即 AmI。这是一个为了解决欧盟日益严重的老龄化带来的养老和医学护理需求所提出的、研究计划规模巨大的信息领域的应用项目。其中包括人们日常活动（Activity of Daily Living, ADL）的识别理解。但目前欧盟的形势变化显然已经缺乏这样的需求。

车辆的自动驾驶，显然是一个重大应用需求的项目。目前卫星定位系统和大数据计算的高度发展。在大多情况下都可以按照已知或固定上下境情况来处理。上下境应如何建模以及如何利用上下境仍然是一个重要的问题。就此而言，与计算机视觉相邻的语音处理和理解、以及物联网 （IoT）中，对上下境的建模和应用方面已有不少成功的经验，可以为我们提供有用的参考。

参考文献

[Dvo 21] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. On the Importance of Visual Context for Data Augmentation in Scene Understanding, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.

[Li 21] Yiming Li, Changhong Fu, Ziyuan Huang, Yinqiang Zhang, and Jia Pan. Intermittent Contextual Learning for Key Filter Aware UAV Object Tracking Using Deep Convolutional Feature. IEEE Transactions on Multimedia, vol. 23, pp. 810-822, 2021.

[Mot 16] Roozbeh Mottaghi, Sanja Fidler, Alan Yuille, Raquel Urtasun, and Devi Parikh. Human-Machine CRFs for Identifying Bottlenecks in Scene Understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 1, pp. 74-87, 2016.

[Rob 65] Larence Gilman Roberts. Machine Perception of Three-Dimensional Solids. In Optical and Electro-Optical, M.I.T Press, pp.159-197, 1965.

[Ser 19] Ygor Rebouças Serpa, Leonardo Augusto Pires, and Maria Andreia Formico Rodrigues, Milestones and New Frontiers in Deep Learning. 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), October, 2019.

[Sze 15] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 1–9, 2015.

[Wei 21] Philippe Weinzaepfel, and Grégory Rogez. Mimetics: Towards Understanding Human Actions out of Context. International Journal of Computer Vision, vol. 129, pp. 1675-1690, 2021.

[Zha 16] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 770–778, 2016.

[威 07] 威根（Wigan, Mark）.《视觉思维》. 孙楠 张伟（译）, 大连理工大学出版社, 2007.

[徐 14] 徐光祐, 邸慧军, 陶霖密. 《人机交互中的体态语言理解》. 电子工业出版社, ISBN 978-121-23625-9, 2014.

责任编委 贾熹滨 张军平 明悦

徐光祐

徐光祐教授 1963 年毕业于清华大学自动控制系，1982 年 12 月-1984 年 12 月美国 Purdue 大学访问学者，1993 年 11 月-1994 年 6 月美国 Illinois 大学访问教授。曾任清华大学计算机系责任教授、博士生导师、IEEE 高级会员、国际测量学会 IMEKO、TC-10 中国代表、中国图象图形学会多媒体技术委员会主席、中国图象图形学报副主编。负责和完成多项 863、科技攻关等重要的国家科研任务，其中包括自然科学重点基金项目“分布式多媒体信息处理方法学及支撑平台研究”（1993－1995，评为优）和 211 重点项目“分布式人机交互”（1997-2000），均取得优秀成果。徐光祐教授率先在国内倡导普适计算的研究，带领清华大学计算机系的部分教师开展了卓有成效的研究工作，被国内外同行认可和推崇。主要表现在：提出“普适计算”这一术语并被广泛接受；国家普适计算相关研究课题的积极建议者，在自然科学基金、863、攻关等国家级项目中均有立项；曾任国际普适计算学术会议（Ubicomp）的中方程序委员；在国家基金委的支持下，组织国内系列普适计算研讨会；主持开展了多项重要的普适计算研究计划，取得显著进展，建立了先进的综合研究基地；与国外同行建立长期稳定的学术联系，力求我们的普适计算研究与国外保持同步；培养出博士生数十名，他们在国内外有影响的学术机构开展相关课题研究；组织形成了来自国内外研究机构、富有创新思维和扎实科学作风的团队。2019 年，徐光祐教授获得 2019 年度中国计算机学会计算机视觉专委会（CCF-CV）终身学术贡献奖。

COMPUTER VISION NEWSLETTER

03 2021
总第 29 期

计算机视觉专委会简报

CCF 计算机视觉专委会