顶会观察

ICCV 2021

微软亚洲研究院视觉计算组研究员 元玉慧

国际计算机视觉大会（International Conference on Computer Vision）是计算机视觉领域的顶级学术会议，与 CVPR 和 ECCV 并称为计算机视觉领域三大顶会。ICCV 属于中国计算机学会推荐国际学术会议中的人工智能领域 A 类会议。今年大会的主席成员包括：来自 Facebook 的 Tamara Berg 和 Tal Hassner、来自麦吉尔大学的 James Clark、来自大阪大学的 Yasuyuki Matsushita、来自宾夕法尼亚大学的 Camillo Jose Taylor、来自蒙特利尔工程学院的 Christopher Pal、来自东京大学的 Yoichi Sato、来自布里斯托大学的 Dima Damen。今年大会最大的研究热点非 Transformer 莫属，在被接收论文中，题目中带有 Transformer 的就有 59 篇，并且最佳论文 Swin Transformer 也是关于 Transformer 的。从被收录论文第一作者的单位来看，43%来自中国，23%来自美国。可以看出，中国在计算机视觉领域的国际舞台上有着举足轻重的地位。

ICCV 2021 于美东部夏令时间的 2021 年 10 月 11 日至 17 日举行。由于新冠疫情的原因，原本安排在加拿大第二大城市蒙特利尔举办的会议转移到了线上平台举办。在会议的线上平台上，参会者可以提前收藏关注自己想要参加的活动或者报告，并且可以直接通过平台提供的 Zoom 会议链接进入直播房间参与讨论。此外，线上平台也支持查询其他参会者的信息、被接收论文的题目和补充材料、获奖论文信息等。大会的第一天和最后两天是以 workshop 和 tutorial 为主，线上平台也提供了 workshop 和 tutorial 的主界面入口。主会是在 10 月 12 日到 15 日之间四天举办的。在主会举办期间，参会者可以自由地与被接收的各篇论文的作者进行面对面的视频交流。下面本文分别从会议概况、论文录用情况、获奖论文的研究工作介绍和精彩的专家观点分享这四个方面进行详细地介绍。

一、 会议概况

James Clark 代表 ICCV 2021 的主席团成员(general chairs)致辞欢迎所有参会者并介绍了会议的安排：82 场 workshops、12 场 tutorials、20 多场企业展览、37 场 mentorship 会议、7 场 affinity 小组会议等。大会最大的三家赞助商包括：谷歌研究院、索尼公司、摩根斯坦利公司(据说这是摩根斯坦利第一次赞助 ICCV 会议)。根据介绍，今年有 4000 多名线上参会者注册了 ICCV 2021。大会还安排了几个特殊环节：例如，两位设计家和艺术家给了一个关于技术发展与社会之间关系的 Keynote Lecture，多位知名计算机视觉领域专家参与了一个关于“计算机视觉中的深度学习方法和传统方法”的论坛等。

二、 论文录用情况

大会的程序主席们(program chairs)对 ICCV 2021 论文的投稿和收录情况作了详细介绍： ICCV 2021 收到了 6152 篇有效投稿，相比于 ICCV 2019 增加了 1800 篇。其中有 11%的投稿选择了撤稿，因此最终有 5486 篇投稿被审稿人评审。大会论文的接受率是 26%，即收录了 1621 篇论文，其中 210 篇论文被选做 oral，另外 1412 篇论文选做 poster。今年会议组织方邀请了 233 位专家作为领域主席(area chair)，平均每位领域主席需要负责 27 篇投稿。组织方邀请了 4216 位

从业者作为审稿人参与论文评审，其中有 2746 位经验丰富的审稿人且每人被分配 7 篇论文、1462 位学生审稿人且每人被分配 4 篇论文、622 位紧急审稿人。每篇论文会收到最少三个评审意见且最终由两位领域主席一起讨论决定是否予以接收。

在被接收论文中，数量最多的研究领域包括：transfer/low-shot/unsupervised learning, image and video synthesis, recognition and classification, detection and localization in 2D and 3D 等。这四个研究领域都有超过 75 篇被录用的论文，其中关于 transfer/low-shot/unsupervised learning 的论文录用数目接近 125 篇。数量最少的研究领域包括 visual reason and logical representation, biometrics, faces 等。大会也按照不同的研究领域分别统计了接收率，其中接收率最高的研究领域包括 gestures and body pose 和 vision for robotics and autonomous vehicles。整体上不同领域的论文接收率区别并不大。

三、 获奖论文选介

大会程序主席 Dima Damen 宣布了 ICCV 2021 的颁奖信息，首先宣布了获得最佳审稿人的名单，从经验丰富的审稿人和学生审稿人中分别选取了前 5%共 210 位作为最佳审稿人。然后宣布了今年的马尔奖(Marr prize)的评委成员并宣布了获奖论文信息：有四篇论文荣获最佳论文提名奖、有一篇论文荣获最佳学生论文奖、有一篇论文荣获最佳论文奖。

最佳论文提名奖：

1 Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. 这篇论文的第一作者 Jonathan T. Barro 来自谷歌研究院，曾因工作 NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis 荣获过 ECCV 2020 最佳论文提名奖。 Mip-NeRF 是为了解决原始 NeRF 在处理不同分辨率的图片时会碰到渲染效果模糊的问题。即使与在不同分辨率的图片上训练的 NeRF 模型相比，Mip-NeRF 的渲染效果也要好很多。不同于 NeRF 采用一个离散的(可阻挡且可发光的)粒子构成连续的体积场(continuous volumetric field)来表示要渲染的场景，Mip-NeRF 提出在一个连续尺度的空间中构建体积场来表示要渲染的场景并采用圆锥的截面而非射线的形式来编码采样空间中的点。实验表明 Mip-NeRF 可以显著提高 NeRF 渲染效果的精细程度且比 NeRF 快 7%，另外， Mip-NeRF 在速度提升 22 倍的情况下取得了与暴力上采样的 NeRF 可比的渲染效果。

2 OpenGAN: Open-Set Recognition via Open Data Generation. 这篇论文的第一作者 Shu Kong 来自卡耐基梅隆大学。为了解决如何判断来自开放集合(即真实世界中)的图片是否包含封闭训练集中出现的类别的问题，OpenGAN 提出结合两种传统做法：学习一个二类的判别器来判断图片是来自开放集合还是封闭训练集合，在封闭训练集合上训练一个对抗生成网络模型(GAN)并利用该模型的判别器来判断图片属于开放集合的可能性。OpenGAN 还发现选择合适的判别器网络结构、对抗地生成一些开放集合的训练数据和在封闭集合的分类器网络抽取的特征上训练判别器对提高最后的实验结果都很重要。

3 Viewing Graph Solvability via Cycle Consistency. 这篇论文的第一作者 Federica Arrigoni 来自特伦托大学。这篇论文是关于运动恢复结构(structure-from-motion)的工作，提出一种新的算法来判断一个视角图(viewing graph)是否可解，即给定的视角图能否推断出一组唯一的投影相机参数。论文中分析了已有的根据多张图片进行唯一三维重建的理论条件并为三维重建理论提供了新的基石。

4 Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. (CO3D) 这篇论文的第一作者 Jeremy Reizenstein 来自 Facebook AI Research。类似于图像识别领域中用于图像检测和分割的 COCO 数据集，CO3D 是目前最大的面向真实世界的三维识别与重建数据集。 CO3D 包含有 150 万个样本，其中每一个样本都包含多个视角拍摄的图像以及图片中物体的三维点云。另外，论文中也提出了 NerFormer 网络结构来利用 Transformer 结构实现三维重建，即根据

几张从不同视角拍摄的图片来重构物体的三维结构。

最佳学生论文奖：

Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. 这篇论文的第一作者 Philipp Lindenberger 来自苏黎世联邦理工大学。论文针对运动恢复结构(structure-from-motion)问题中的两个关键步骤进行了改进：首先，调整初始二维关键点位置先验到任意一个几何估计条件下，然后，对三维点云和相机坐标做后处理。这样的改进操作对检测噪声以及外观变化都非常鲁棒。实验验证了该改进方案在各种关键点检测任务中可以提高相机姿态估计和场景几何估计的效果。

最佳论文奖：

Swin Transformer: Hierarchical Vision Transformer using Shifted Window. 这篇论文的共同第一作者 Ze Liu、 Yutong Lin、 Yue Cao、 Han Hu 来自微软亚洲研究院、中国科技大学、西安交通大学。 Swin Transformer 是一种通用网络结构，其核心思想是：利用一个滑动窗机制来实现局部区域之间的信息传播、只在局部窗口区域内应用自注意力从而降低计算复杂度、采用分层次的网络结构设计来学习多尺度的特征表示等。实验结果验证了 Swin Transformer 在图像分类、物体检测和语义分割任务上都取得了比卷积神经网络显著更好的结果，从而验证了 Transformer 在计算机视觉领域的巨大潜力。

大会也宣布了 PAMI-TC 奖的获得者：来自加州伯克利大学的 Ruzena Bajcsy 获得了 2021 Rosenfeld Lifetime Achievement Award，来自加州理工学院的 Pietro Perona 和来自 INRIA 的 Cordelia Schmid 荣获了 2021 PAMI Distinguished Researcher Award，The KITTI Vision Benchmark Suite 团队和 The Detectron object detection and segmentation software 团队荣获了 2021 Everingham prize。最后也对三篇发表于十年前 ICCV 2011 上的工作颁发了 2021 Helmholtz Prize。

1 ORB: An efficient alternative to SIFT or SURF, ICCV 2011. 这篇论文的第一作者 Ethan Rublee 来自机器人实验室 Willow Garage。 ORB 在 2011 年论文发表的时候是一种高效的视觉任务描述子(descriptor)且比 SIFT 的速度快 2 个数量级。实验结果验证了 ORB 在当时移动端(运行在手机上)的物体检测和区块跟踪等任务上都表现很好。

2 HMDB: A large video database for human motion recognition, ICCV-2011. 这篇论文的第一作者 Hildegard Kuehne 来自德国的卡尔斯鲁厄理工学院(KIT)且目前在德国的法兰克福大学工作。这篇论文的贡献是构建了在当时最大的用于动作识别的视频数据集 HMDB, HMDB 数据集包括接近 7000 段视频片段，定义了 51 个不同的语义类别标签。

3 DTAM: Dense tracking and mapping in real-time, ICCV 2021. 这篇论文的第一作者 Richard Newcombe 来自帝国理工大学，目前 Facebook Reality Labs 工作。DTAM 在当时是一个用于实时相机跟踪和三维重建的算法。DTAM 不依赖于当时基于复杂特征提取算法的特征匹配方法，而是直接利用 RGB 信息对图像原始像素做匹配。

四、 精彩报告选介

大会内容精彩纷呈，由于篇幅受限，这里仅仅选取其中最有代表性的几位计算机视觉专家在 “A discussion about deep learning vs classical methods and their roles in computer vision” 论坛上的精彩分享为例作详细地介绍。

来自华盛顿大学的 Richard Szeliski 分享了他编写的最新版的教材 Computer Vision: Algorithms and Applications(第二版)中关于会议主题的讨论。另外，Richard Szeliski 也分享了深度学习方法已经被华盛顿大学、密歇根大学、麻省理工学院的计算机视觉课程所囊括。然后， Richard Szeliski 选取了多篇关于应用深度学习方法解决传统计算机视觉任务的工作，比如三维重建、深度估计、相机参数估计、光流估计、图像合成等。最后， Richard Szeliski 给出了他对这个问题的建议：深度学习仅仅是很多有效工具中的一种方法但

是不是唯一方法、深度学习并不适合依赖几何/光学/物理约束的问题和建模不确定性的问题、深度学习方法在训练数据不够的情况下很难具有较好的泛化能力、基于深度学习的特征提取速度在有些任务上比传统方法快很多。

来自加州伯克利大学的 Jitendra Malik 从 1973 年的诺贝尔生理学奖获得者 Nikolaas Tinbergen 提出的关于研究动物行为的四个经典问题(注解：(1) adaptive function and (2) phylogenetic history; and the proximate explanations, in particular the (3) underlying physiological mechanisms and (4) ontogenetic developmental history https://en.wikipedia.org/wiki/Tinbergen's_four_questions)出发，用人类如何估计深度作为例子来解释计算机视觉任务。 Jitendra Malik 认为未来应该研究如何设计支持小数据规模而非大数据规模、利用自然信号而非人工标注作为监督信号的算法。最后， Jitendra Malik 也鼓励计算机视觉领域的研究人员去学习人类思想史并分享了自己如何汲取计算机视觉领域的思想史来做出更好研究工作的经验。

来自伊利诺伊大学厄巴纳-香槟分校的 Svetlana Lazebnik 从回顾 AlexNet 在 ECCV 2012 ImageNet 挑战上取得了非常好的分类结果开始，提到了当时她和周围的研究员最开始对深度学习方法的推广(因为 GPUs 不普及)保持怀疑态度。随着更方便的深度学习框架 Caffe 的出现，在 2013 年、2014 年、2015 年发表深度学习方面的顶会论文非常容易。例如，当时只需要训练一个 AlexNet 或者把 AlexNet 抽取的特征用于各种各样的计算机视觉任务就可以发一篇顶会论文。现在由于很多 low-hanging fruits 快都被摘完了，所以深度学习研究也进入了一个瓶颈期。最近随着 GPT 和 Transformer 的成功应用，Svetlana Lazebnik 也表示她对深度学习发展没有之前那么悲观。

来自加州伯克利大学的 Alexei Efros 回顾了自己早在 2012 年就访问纽约大学研究深度学习的计算机视觉实验室。Alexei Efros 也坦言自己不喜欢追随热点去研究网络结构，而是喜欢思考计算机视觉问题的本质。Alexei Efros 也回顾了早在 2015 年自己就带领团队开始专注于自监督学习并发表了多篇顶级论文。

五、 总结与展望

通过深度参与今年的 ICCV 大会，我们不仅关注到 Transformer 快速席卷各种计算机视觉领域各个重要任务并有替代卷积神经网络的趋势，也关注到最前沿的专家们分享了关于计算机视觉领域的研究工作目前面临的重要挑战与局限性，同时还关注到专家们对深度学习方法和传统方法之间关系的热烈讨论，以及工业界专家们分享如何将计算机视觉领域的前沿技术落地到自动驾驶系统中等。笔者相信未来计算机视觉领域的研究热点应该会包括：如何推动计算机视觉领域的大模型预训练、如何构建视觉大模型预训练所需要的大规模高质量的数据集、如何提高算法模型在开放世界中的泛化能力等。笔者认为回答好这些问题将大大推动计算机视觉向更通用的智能迈进。

责任编委 **魏秀参**

元玉慧

微软亚洲研究院视觉计算组研究员。主要研究方向为语义分割和物体检测。
Email: yuhui.yuan@microsoft.com