专题综述

关于构建遥感图像解译数据集的探讨

武汉大学 夏桂松 龙洋

一、引言

近年来，遥感图像解译技术发展迅速，大幅提升了人类对地物信息的感知能力。随着遥感图像获取能力的不断增强，积累的遥感图像数据越来越多，实际应用中对海量遥感图像的自动化、智能化解译需求也越来越大，而遥感图像基准数据集是研发和测试相关解译算法的前提。因此，如何构建可靠的遥感图像解译基准数据集，为训练、测试和筛选实用的解译算法提供数据支撑，是推动遥感图像自动化、智能化解译发展的关键之一。

通过遥感图像解译算法提取图像中有价值的信息，是实现遥感图像理解和应用的基础。然而遥感图像复杂的光谱和结构等属性特征，为遥感图像内容的解译带来了严峻挑战。近年来，以深度学习为代表的数据驱动方法已经成为人工解译的重要替代方法，在实现大规模遥感图像自动解译和内容理解方面展现出巨大的潜力。然而，由于遥感图像解译标准数据集的缺乏，遥感图像解译算法的发展仍然面临着诸多挑战：接收的遥感图像数据量越来越大，但其中大部分数据未被赋予有价值的标注信息，数据难以被有效利用；缺乏有代表性的、具有精确语义标注信息的大规模遥感图像数据集，遥感图像解译算法的发展因此受到极大限制；遥感图像解译算法的泛化能力不足，难以满足复杂现实场景的应用需求；缺乏可靠的算法公共测试平台，难以对不同的解译算法进行系统评价和公平比较。

因此，建立可靠的遥感图像数据库构建原则与方法，构建面向实际应用场景的大规模遥感图像标注数据库，发展遥感图像解译算法测试与评价的公共平台，进而推动遥感图像解译算法面向实用化、自动化方向发展，是

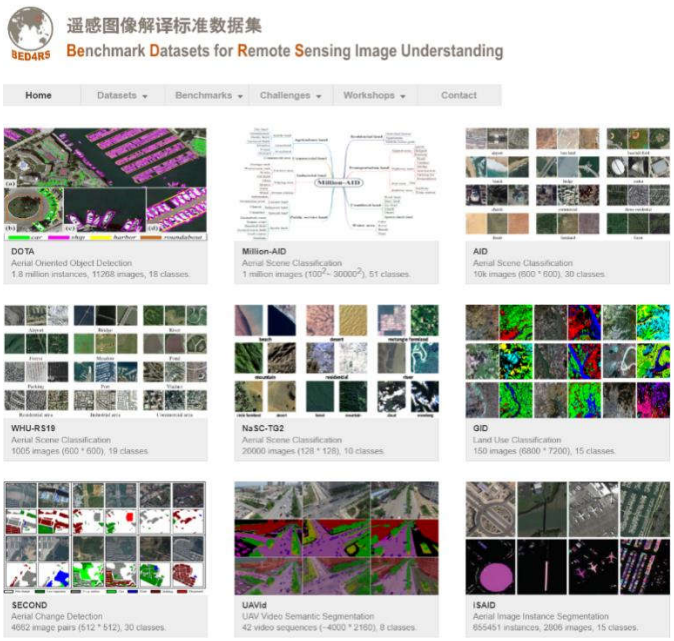

图 1 大规模遥感图像解译数据库

应对上述挑战的一种有效途径。研究团队目前已构建了面向遥感图像智能解译的大规模标准数据库，包括面向遥感图像场景分类、目标检测、语义分割和变化检测等任务的精确标注数据集，如图 1 所示。

二、遥感图像解译数据集构建准则

数据驱动的遥感图像解译算法的性能严重依赖于训练数据集中语义标注信息的规模和质量。构建高大规模、高质量图像解译数据集的主要挑战在于数据集构建的效率和质量控制。本文探讨构建遥感图像解译数据集的基本准则和方法，以期为大规模和实用化遥感图像解译数据集的高效构建提供参考指导。

一方面，遥感图像解译数据集的构建应该面向实际

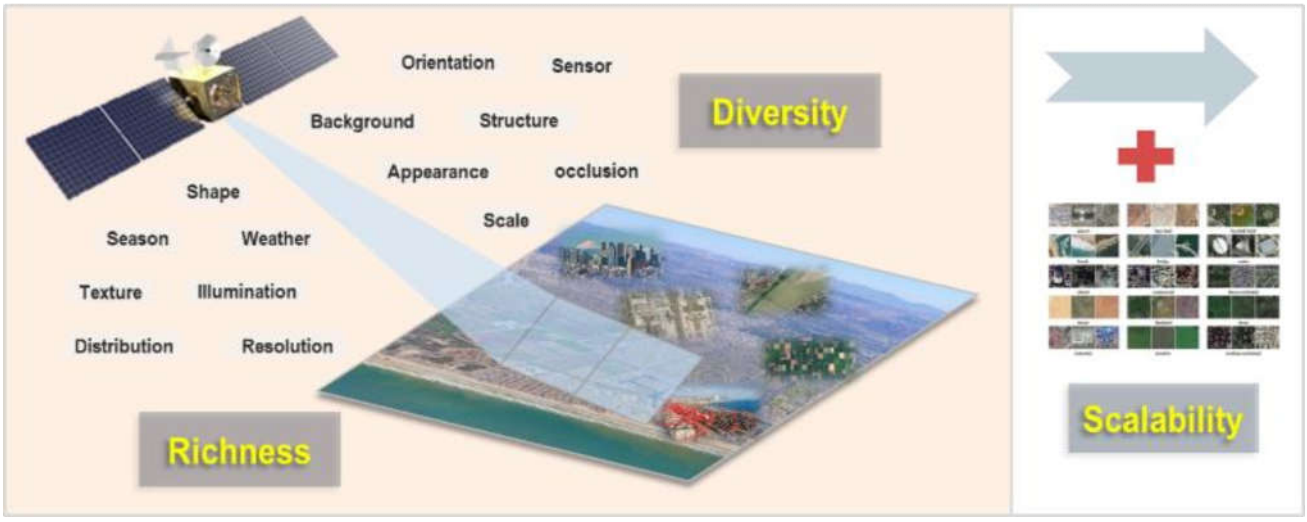

图 2 遥感图像解译数据集构建基本原则：多样性（Diversity），丰富度（Richness），和扩展性（Scalability）

应用需求，而非面向解译算法的特性。事实上，遥感图像解译数据集应该针对实际应用中解译算法的训练、测试和筛选而构建。因此，可靠的基准数据集对于全面评估所设计的解译算法的性能，进而挑选出实用的解译算法至关重要。另一方面，遥感图像解译数据集的标注应由应用部门而非由算法开发人员来完成。算法开发人员对算法属性比较熟悉，在标注过程中会不可避免地带来个人偏好，从而导致所标注的解译数据集偏向于算法特性。相对而言，应用人员对遥感图像解译的现实应用场景有更深入的理解，因此更熟悉解译任务中存在的具体挑战。因此，后者标注的数据集更有利于增强算法的实用性。基于以上认识，本文提出基于多样性（Diversity），丰富度（Richness）和可扩展性（Scalability）的遥感图像解译数据集构建准则，如图 2 所示。

2.1. 多样性

数据集中的图像样本能够反映感兴趣地物在光谱、几何、形状、纹理等方面的属性特征，可认为该数据集具有良好的多样性。从类内多样性的角度来看，数据集中的每个标注样本应能从不同方面反映同一类别地物的不同属性特征，而非地物内容和图像样本的简单重复。因此，多样性较强的标注样本能够更全面地表征感兴趣地物在现实世界中的分布模式，进而为训练具有更强特征表达能力的遥感图像解译算法提供可靠保障。

另一方面，在构建遥感图像解译数据集时还应考虑不同类别地物之间的相似性。为此，遥感图像数据集中可以包含具有高度语义重叠和相似特征的细粒度类别，从而使解译模型学习到区分不同类别地物的本质特征。增强感兴趣地物类内多样性和类间相似性，是丰富地物特征的一种有效途径，这对于构建具有较强多样性的遥感图像解译数据集至关重要，进而增强所构建的解译数据集的实用性。

2.2. 丰富度

除了多样性，遥感图像解译数据集的丰富度也是学习具有较强稳定性解译算法的重要保障。具体来说，遥感图像解译数据集，应具有丰富的图像样本、包含丰富的地物信息。为此，构建遥感图解译数据集时，需要采集不同天气、不同季节、不同光照、不同成像条件、不同传感器、不同时间和空间等条件下的遥感图像，体现感兴趣地物在平移、视角、对象姿态和外观、空间分辨率、光照、背景、遮挡、时空属性等方面的特征差异。

此外，遥感图像以俯视视角拍摄，具有地理覆盖范围大、包含地物丰富、背景信息复杂等特点。面对这种情况，解译数据应包含具有多样化特征的图像场景，如在几何形状、结构特征、纹理属性等方面的多样性。从这一角度来看，数据集应该包含大尺度的图像和足够多的标注样本，以体现地物特征的分布模式。现实中经常会由于图像和样本的不足导致解译模型学习出现过拟合的现象，这一问题在数据驱动的解译算法(如卷积神经网络)中尤为显著。因此，基于以上考虑建立具有较强丰富度的遥感图像解译数据集，能够使得所构建的解译模

型具有更强的泛化能力。

2.3. 可扩展性

可扩展性可用于描述已构建的解译数据集的扩展应用能力。遥感图像解译应用日益广泛，现实中对解译数据集的需求通常会随着时间和应用场景的变化而变化。例如，随着土地覆盖和土地利用的变化，可能需要将新的土地利用类别与数据集已构建的类别体系区分开来。因此，构建的解译数据集需要有充足的类别扩展空间，以包含新的地物类别，同时对不同地物类别之间的关系进行有效组织。因此，考虑到实际应用场景和应用需求的变化，解译数据集需要具有较好的可扩展性。

值得注意的是，现实中每天接收的遥感图像类型多样且规模巨大，迫切需要采用高效的方式为其赋予标注信息，从而发挥应用价值。因此，对遥感图像及其标注信息进行合理的组织、保存和维护对于数据集的可扩展性同样具有重要意义。例如能够将新标注的图像无缝地集成于已构建的数据集中，是数据集应用具有可扩展性的重要体现。因此具有良好可扩展性的解译数据集可以有效适应现实应用场景的需求变化。

三、遥感图像解译数据库构建方法

3.1. 数据库图像语义标注

根据标注过程是否有人工参与以及人工参与的程度，遥感图像数据集的语义标注方法可分为三种类型，即人工标注、自动标注和交互式标注。

人工标注：人工标注的过程是一种完全监督的标注模式，其优势在于具有较高的标注精度，因而许多遥感图像数据集采用人工标注的方式构建。无论是自然图像还是遥感图像，对图像中的内容进行标注的方式都是相似的，并且为了提高图像标注效率，现阶段已开发了许多面向不同解译任务的标注工具。因此，针对自然图像开发的图像标注工具可以进一步引入到遥感图像中，为构建高质量的遥感图像解译数据集提供基础。

自动化标注：遥感图像所包含的地物内容复杂，对于缺乏领域知识的标注者来说，很难对其语义内容进行精确标注。此外，手工标注的方式容易因标注者在领域知识和标注技能等方面的差异产生偏差，因而采用自动化标注的方法可以降低手工标注的难度，并进一步提高标注效率。自动化标注的方法通过构建一定数量的初始化样本来训练一个解译模型，然后将待标注样本输入到建立的标注模型中进行内容解译，最终将解译结果作为标注信息。由于遥感图像具有地理范围大、内容复杂的特点，可以采用迭代学习或增量学习来过滤噪声标注，提高标注模型的泛化能力。

交互式标注：在遥感大数据时代，现实应用中对遥感图像标注质量和效率需求不断提高，基于人机交互的半自动化标注是一种更为实用的标注方案。在该方案中，可以利用已有的标注数据构建初始标注模型，然后对未标注的遥感图像进行自动化标注。通常，通过使用主动学习策略和设置约束条件，利用解译模型筛选出难标注的图像，再采用人工进行标注，并将人工标注的信息反馈给自动标注模型，最后通过迭代学习的方式来进一步优化标注模型的性能。随着交互标注和迭代学习过程的进行，需要标注的图像数量将大大减少，从而大幅减轻标注难度并减少标注工作量。因此交互式标注的整体性能主要取决于标注者参与交互式标注的时间。

3.2. 图像语义标注质量控制

高质量的遥感图像数据集标注信息对于开发有效的解译算法及其性能评价十分重要，因此需要采用可靠的策略来控制数据集图像标注质量。

规则和样本：建立明确的标注规则是构建高质量遥感图像解译数据集的基础，否则不同的标注者将根据各自认知和偏好进行语义信息标注，从而对标注质量和标注信息的标准化管理造成影响。对于遥感图像标注而言，需要由具有领域知识的图像解译专家，建立合理的标注规则和示例，以建立良好的数据标注基础。

标注人员培训：通过对标注人员的培训，提高标注队伍的专业素质，进而为数据集标注质量提供保证。具体地，可以为每个标注员分配待标注数据，并要求其按规则进行标注，最后剔除未能通过测试评估的标注员。

多阶段标注：一系列复杂的标注操作容易引起标注者的疲劳并导致错误标注。为消除这种影响，可以设计多阶段的图像标注策略，以减小复杂标注任务的难度。通过这种处理，每位标注员只需关注整个标注工程中的

一个简单步骤，从而有效地降低标注错误率。

多重标注：采用多个标注者对同一个对象进行标注，并将不同的标注结果合并，可以有效提升数据集的标注精度。然而该方法的不足之处在于一个标注对象需要多个标注人员进行重复标注，因而标注效率较低。

标注审查：邀请相关标注者开展同行评审并对标注结果的质量进行评级，还可由领域专家进行进一步的评审。通过对标注流程中不同层次和步骤的标注结果进行严格的监督审查，可以实现对整体标注结果的质量控制。

四、大规模遥感图像解译标准数据库

为了促进遥感图像解译算法的研究和发展，研究团队基于所提出的遥感图像解译数据集构建准则和方法，建立了大规模遥感图像解译标准数据库，包含面向场景分类、目标检测、语义分割和变化检测等不同解译任务的大规模遥感图像标注数据集，以期为相关研究人员提供可靠的数据基础和解译算法测试标准。数据库访问地址：https://captain-whu.github.io/BED4RS

4.1. 遥感图像场景分类数据集：AID/Million-AID

AID 数据集包含 30 个场景类别，共 1 万幅场景实例。在 AID 基础上，研究团队建立了半自动化的遥感图像场景标注方案，并构建了百万级遥感图像场景分类数据集 Million-AID，共包含 51 个土地利用类别。与已有的遥感图像场景分类数据集相比，研究团队构建的场景分类数据集地理分布范围更广、场景样本丰富、数据集规模大。不同的场景类别采用层次化的类别网络进行语义关系组织，每个语义类别包含 2000~45000 幅场景图像，能满足数据驱动的解译模型构建与优化需求。

4.2. 遥感图像目标检测数据集：DOTA

DOTA 是一个大规模遥感影像目标检测数据集，目前有 DOTA-V1.0/V1.5/V2.0 三个版本。DOTA 数据集包含来自 GF-2、JL-1 和 Google Earth 等卫星和平台的高分辨遥感图像，图像尺寸变化大（800~13000），目标类别丰富，标注样本规模大。每一个目标对象采用四边形边界框进行标注，可以精确表征目标的位置和方向。DOTA-V1.0 包含 2806 幅遥感图像，涵盖 15 个常见对象的类别共 188282 个标注实例；DOTA-V1.5 在 DOTA-V1.0 基础上进行扩展，包含 403318 个标注实例；DOTA-V2.0 包含 11268 幅遥感图像，涵盖 18 个语义类别共 1793658 个标注实例。此外，研究团队为 DOTA 数据集提供了开放的算法测试与性能评估平台。

4.3. 遥感图像语义分割数据集：GID

GID 是一个用于遥感图像语义分割的大规模数据集，包含覆盖中国 60 多个不同城市的高分二号卫星图像，覆盖面积超过 50000 平方公里。GID 数据集由两部分组成：广域分类集和精细分类集。广域分类集包含 5 个语义类别共 150 景像素级标注的高分二号图像，精细分类集包含 15 类共 3 万个场景级标注样本和 150 景像素级标注的高分二号图像。GID 具有覆盖范围大、分布区域广和空间分辨率高等优点，相比现有的土地覆盖分类数据集能更好地满足广域制图的应用需求。

4.4. 遥感图像变化检测数据集：SECOND

大规模语义变化检测数据集 SECOND 包含 4662 对大小为 512x512 的遥感图像。各图像主要分布于成都、上海以及杭州等城市。所有图像像素都被精确地标记为 30 类语义变化类别，即低矮植被、树木、无植被土地、水体、建筑和运动场等地物类别之间的变化。与传统的语义变化检测数据集不同，SECOND 包含同一地物类别之间的变化(建筑物的拆除重建)，可以为研究遥感地物语义类别之间的变化提供更好的数据基础。

五、 未来发展方向展望

现阶段，研究团队所构建的上述遥感图像解译数据集已经集成到业界最大遥感影像样本库 LuoJiaSET 中，并成功应用于全球首个遥感影像智能解译专用深度学习框架 LuoJiaNET 的构建。在未来的工作中，研究团队将致力于构建面向遥感图像解译的数据集和解译算法在线公共测试平台，并建立遥感图像智能解译算法库。我们相信，遥感图像解译智能化、自动化的发展趋势不可阻挡，未来将会有更多面向真实遥感应用场景的数据集和算法不断涌现，因此应该鼓励更多的遥感图像解译数据集和解译框架在社区内共享，以促进遥感图像智能解译和应用的发展繁荣。

责任编委 储珺

参考文献

[1] C. Toth and G. Jóźków. Remote sensing platforms and sensors: A survey. ISPRS J. Photogrammetry Remote Sens., 2016, 115: 22–36.

[2] T.-Z. Xiang, G.-S. Xia, and L. Zhang. Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects. IEEE Geosci. Remote Sen. Mag., 2019, 7(3): 29–63.

[3] G.-S. Xia, J. Hu, F. Hu, et al. AID: A benchmark data set for performance evaluation of aerial scene classification. IEEE Trans. Geosci. Remote Sens., 2017, 55(7): 3965–3981.

[4] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. Proc. IEEE, 2017, 105(10): 1865–1883.

[5] Y. Long, G.-S. Xia, S. Li, et al. On creating benchmark dataset for aerial image interpretation: reviews, guidances and Million-AID. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., 2021, 14: 4205-4230.

[6] G.-S. Xia, X. Bai, J. Ding, et al. DOTA: A large-scale dataset for object detection in aerial images. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018: 3974–3983.

[7] J. Ding, N. Xue, G.-S. Xia, et al. Object detection in aerial images: a large-scale benchmark and challenges. IEEE Trans. Pattern Anal. Mach. Intell., 2022: 1-18.

[8] M. D. Hossain and D. Chen. Segmentation for object-based image analysis: A review of algorithms and challenges from remote sensing perspective. ISPRS J. Photogrammetry Remote Sens., 2019, 150: 115–134.

[9] X.-Y. Tong, G.-S. Xia, Q. Lu, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models. Remote Sens. Environ. 2020, 237: 111322.

[10] K. Yang, X.-Y. Tong, G.-S. Xia, et al. Hidden path selection network for semantic segmentation of remote sensing images. IEEE Trans. Geosci. Remote Sens., 2022: 1-13.

[11] X. Zhu, C. Vondrick, C. C. Fowlkes, et al. Do we need more training data? Int. J. Comput. Vis., 2016, 119(1): 76–92.

[12] K. Yang., G.-S. Xia, Z. Liu, et al, Asymmetric siamese networks for semantic change detection in aerial images. IEEE Trans. Geosci. Remote Sens., 2022, 60: 1-18.

[13] D. Acuna, H. Ling, A. Kar, et al. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018: 859-868.

夏桂松

武汉大学计算机学院教授、博士生导师。研究方向:计算机视觉、模式识别、机器学习及应用。
主页：http://www.captain-whu.com/xia.html
Email: guisong.xia@whu.edu.cn

龙洋

武汉大学测绘遥感信息工程国家重点实验室在读博士生。主要研究方向为遥感图像理解。
Email: longyang@whu.edu.cn

专题综述

对抗学习：消除对抗噪音以提高对抗鲁棒性

西安电子科技大学 周大为 王楠楠

随着人工智能领域的快速发展，深度神经网络得到了广泛的应用。但是，最近研究发现深度神经网络对不易察觉但具有对抗性的微小扰动（即对抗噪音）具有明显的脆弱性。为了缓解对抗噪音的影响，对抗防御的研究受到了越来越多的关注。基于预处理的对抗防御是主要的防御类别之一，这种方法期望通过消除对抗噪音来提高深度神经网络的对抗鲁棒性。然而，基于预处理的防御可能会受到扰动放大效应的影响。此外，生成对抗噪音的对抗攻击算法是在不断演变的，基于已知类型的对抗噪音训练的防御模型通常不能很好地泛化到未知类型的对抗噪音上。为了解决这些问题，使用高层的类激活特征来消除对抗噪音和通过学习攻击不变表征来消除对抗噪音是值得探索的基于预处理的对抗防御策略。

一、神经网络对抗噪音的脆弱性

人工智能处在现代数据驱动科学的前沿，机器学习是人工智能的核心部分。深度神经网络是机器学习中重要的部分，其已经被广泛应用于商业、卫生和国防等领域。然而，大多数深度神经网络都存在严重的漏洞。他们很容易被微小的、人类无法察觉但经过精心设计的噪音（即对抗噪音）误导[1][2]。

以分类任务为例，如图 1 所示，在自然样本上添加对抗噪音后生成的对抗样本在人类视觉上与自然样本没有明显的区别，但是他们会使深度神经网络产生错误的预测结果。深度神经网络的这种脆弱性对许多安全敏感型应用（比如人脸识别、自动驾驶）构成了严重风险。

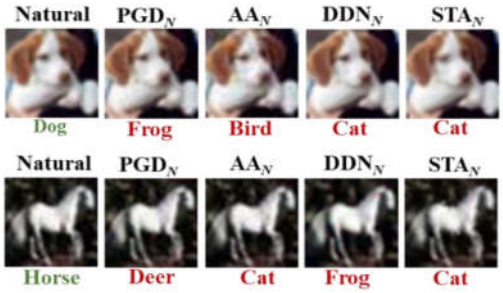

图 1 自然样本（Natural）与不同的对抗样本。两者在人类视觉上相似，但是深度神经网络对对抗样本会做出错误的预测。下标“N”表示相应的攻击是非目标攻击。

为了提升深度神经网络对对抗噪音的鲁棒性（即对抗鲁棒性），针对对抗防御的研究得到了越来越多的关注。基于预处理的对抗防御是主要类别之一。已有研究表明，对输入样本中的对抗噪音进行消除[10]，可以增强目标模型的对抗鲁棒性。但是，如何设计有效的方法缓解扰动放大效应的影响仍然需要进一步研究。此外，对抗攻击算法是多样且在不断发展的，基于已知对抗噪音训练的防御模型如何对未知对抗噪音具有较好的泛化性也值得深入探索。

二、基于类激活特征的对抗防御

扰动放大效应是指微小残留的对抗噪音在深度神经网络内部层中被逐渐放大，并最终导致网络输出错误的预测。类激活映射技术[11]为解决这个问题提供了一种潜在的方法。给定一个分类网络，类激活映射技术可以通过将输出层的类别权重投影回最后的卷积层特征，并对加权特征进行线性求和来识别输入图像每个区域对最终预测的重要性。如图2所示，虽然对抗噪音在像素级是不易察觉的，但自然样本的类激活图和对

抗样本的类激活图之间存在明显差异。此外，加权特征位于网络的高层，在该层中，残余对抗噪音导致了较大的扰动。这促使我们可以在类激活特征空间中设计一种防御方法来抑制扰动放大效应。

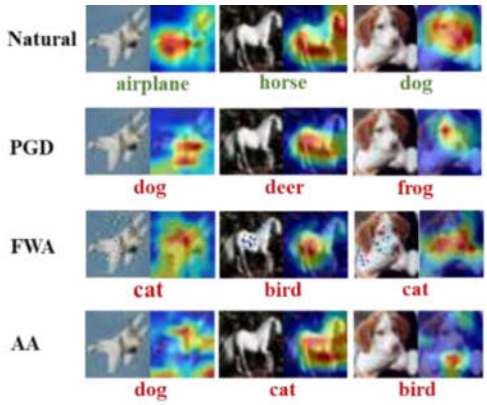

图 2 自然样本和对抗样本的类激活图。虽然对抗噪音在像素级是不易察觉的，但自然样本和对抗样本的类激活图之间存在明显差异。

我们设计了一种通过利用类激活特征来消除对抗噪音的防御方法。这种方法在类激活特征空间中，以自监督的对抗训练方式训练去噪模型，而不需要额外类型的对抗样本和真实的类别标签。具体来说，首先通过最大限度地破坏自然样本的类激活特征来生成对抗噪音并制作对抗样本。类激活特征的差异使得深度神经网络对对抗样本和自然样本做出不同的预测结果。我们将这种攻击称为基于类激活特征的攻击，优化目标为：

$$\max_{\tilde{x}} \Delta(x,\tilde{x}) \quad \text{,subject to: } \|x-\tilde{x}\|_\infty \le \epsilon \ ,$$

其中，$\Delta(x,\tilde{x})=\delta(\Phi_x,\Phi_{\tilde{x}})$，$\delta(\cdot)$ 为2范数度量，Φ_x 和$\Phi_{\tilde{x}}$表示自然样本和对抗样本的类激活特征，ϵ为扰动边界。

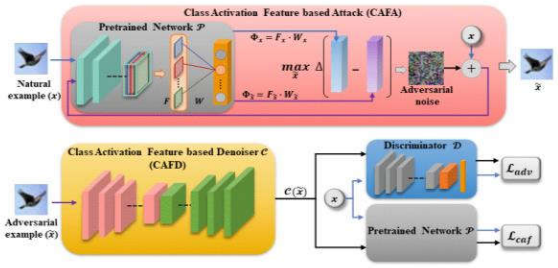

图 3 基于类激活特征的对抗防御方法的框架图。

基于生成的对抗样本，我们训练一个去噪模型，即基于类激活特征的去噪器(CAFD，class activation feature-based denoiser)。此方法没有直接使用像素级损失函数来训练防御模型，而是最小化自然样本和对抗样本的类激活特征之间的距离。另外，引入一种基于RaGAN[12]的图像鉴别器来增强去噪后图像的纹理细节，使其更接近自然样本。方法的整体框架如图3所示。

在SVHN[13]和CIFAR-10[14]两个数据集上的评估能够说明上述方法的有效性。使用三种具有代表性的攻击来生成对抗样本，分别为PGD[3]，AA[4]和FWA[7]。如图4所示，基于类激活特征的对抗防御方法能够消除对抗噪音，恢复对抗样本的类激活图。定量分析的结果如表1所示，相比于之前的JPEG[17]、TVM[17]、APE-G[18]和HGD[19]方法，对抗准确率有所提升。

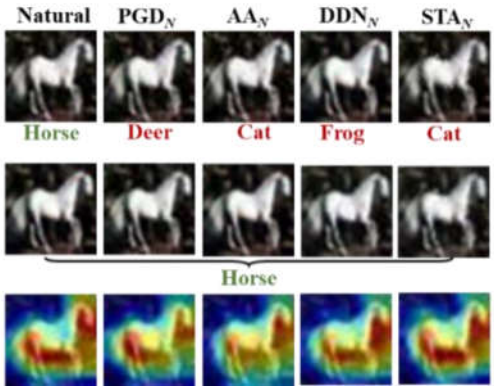

图 4 对抗防御方法针对不同的对抗攻击的去噪效果。(第一行：自然样本和对抗样本；第二行：去噪后的样本；第三行：去噪后的样本的类激活图)。

表 1 防御不同对抗攻击生成的对抗样本的对抗准确率。APE-G 和 HGD 使用 DDN_N 生成的对抗样本作为对抗性训练数据。

Dataset	Defense	NONE	PGD_T	PGD_N	AA_N	FWA_N
SVHN	JPEG	90.22	13.40	4.33	2.44	6.93
	TVM	89.99	24.83	5.47	4.16	3.39
	APE-G	89.60	16.60	6.80	18.08	13.43
	HGD	89.88	55.00	42.65	37.75	32.56
	CAFD	**92.35**	**89.37**	**85.36**	**86.43**	**41.93**
CIFAR-10	JPEG	86.68	49.66	48.59	41.16	14.73
	TVM	90.35	43.29	31.21	33.54	9.190
	APE-G	91.82	37.69	21.92	23.35	23.34
	HGD	**92.36**	68.82	53.13	54.27	53.13
	CAFD	91.10	**89.42**	**87.21**	**88.20**	**64.41**

三、基于攻击不变特征的对抗防御

专注于有限训练数据中已知类型的对抗样本可能导致对抗防御模型过拟合于给定类型的对抗噪音，并缺

乏针对未知类型对抗噪音的通用性或有效性。现实世界中存在广泛甚至未知的对抗攻击类型，这促使我们设计一种可以处理不同类型或未知的对抗样本的防御方法。

认知科学中的一些研究对解决这个问题具有一定的启发。具体来说，一些研究表明[20][21]，即使人脸显示出不同甚至未曾见过的表情，我们也能够识别人脸身份，因为我们的大脑善于提取不变的面部特征。同样，如图 5 所示，我们人类也无法轻易区分自然样本和对抗样本，因为我们只关注表示语义分类信息的不变特征，而忽略了对抗噪音。对抗样本保留了不变特征，以使在人类视觉上无法预先识别对抗样本。我们将这种不变特征命名为攻击不变特征。

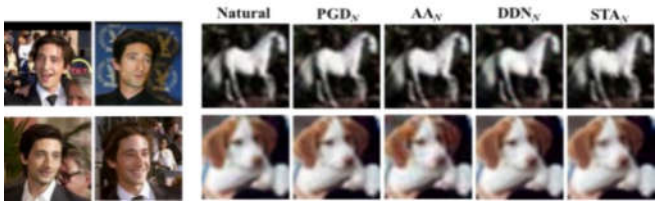

图 5 同一身份不同表情的人脸图像和不同的对抗样本。

基于攻击不变特征，我们设计了一种对抗防御方法(AIFD, attack-invariant feature-based defense)来消除对抗噪音。这种方法基于自动编码器的框架，将对抗噪音消除分为学习攻击不变特征和从攻击不变特征中恢复自然样本。方法的整体框架如图 6 所示。

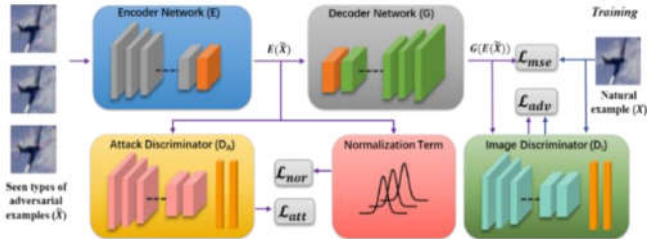

图 6 基于攻击不变特征的对抗防御整体框架图。

具体来说，我们以对抗性特征学习的方式引入了一对编码器和鉴别器，用于将攻击不变特征从对抗样本中分离出来。鉴别器用于从编码的攻击不变特征空间中区分特定于攻击的信息(例如，攻击类型标签)，而编码器旨在提取出使鉴别器无法区分的特征。通过迭代优化，在编码的过程中可以去除特定于攻击的信息，并保留不变特征。编码器和鉴别器对应的损失函数分别是：

$$\mathcal{L}_{D_A} = -\frac{1}{K}\sum_{k=1}^{K} Y_k^P \cdot log(\sigma(D_A(E(\tilde{X}_k)))) \ ,$$

$$\mathcal{L}_{att} = -\frac{1}{K}\sum_{k=1}^{K} Y_\zeta^P \cdot log(\sigma(D_A(E(\tilde{X}_k)))) \ ,$$

其中，σ(·)表示 softmax函数，$\tilde{X}_k$表示第k种对抗攻击生成的对抗样本，$Y_k^p = [y_{k1}^p, y_{k1}^p, \dots, y_{kN}^p]^T$ 表示第 k 种攻击的攻击特定标签，$y_{kn}^p = [\xi_1, \xi_2, \dots, \xi_K]^T$表示 one-hot 向量且$\xi_i$在$i = k$时等于 1，其他时候为 0。$Y_\zeta^p = [y_{\zeta 1}^p, y_{\zeta 2}^p, \dots, y_{\zeta N}^p]^T$ 是攻击混淆标签，$y_{\zeta n}^p = [1/K, 1/K, \dots, 1/K]^T$是一个 k 维常数向量。K 为使用的对抗攻击的类别数。

另外，训练过程中使用的对抗样本往往带有偏差，因为现实世界中广泛存在的攻击类型非常多样。偏差问题可能会使学习到的编码器适用于训练过程中使用的攻击或类似类型的攻击，但对某些明显不同的未知类型的攻击具有较差的泛化能力。为了解决偏差问题，我们在攻击不变特征的编码空间中添加一归一化项，以将每种类型的攻击的特征分布与多元高斯先验分布[15][16]相匹配。通过这种设计，学习到的攻击不变特征有望推广到未知类型的攻击。该项的损失函数为

$$\mathcal{L}_{nor} = JSD(P_1, \dots, P_K) = -\frac{1}{K}\sum_{k=1}^{K} K\mathcal{L}(P_k || \mathcal{N}).$$

此外，为了提高从攻击不变特征中恢复的样本的质量，引入图像鉴别器和像素级距离度量作为约束来帮助训练解码器。通过联合优化编码器和解码器，防御模型可以有效地消除多种类型的对抗噪音，提高对抗准确率。

如图 7 和表 2 所示，在 MNIST[9]和 CIFAR-10 两个数据集上的定性和定量分析表明，针对六种对抗攻击 PGD，AA，FWA，CW[5]，DDN[6]，STA[8]。该方法能够有效去除对抗噪音。相比于之前的方法 APE-G 和 HGD，对抗准确率有所提升。

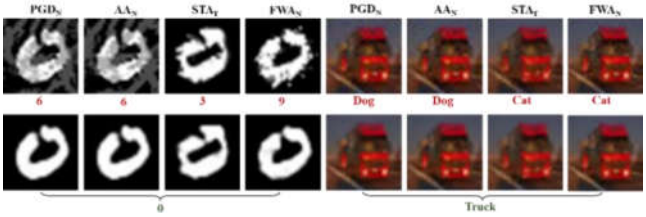

图 7 对抗防御方法针对不同的对抗攻击的去噪效果。(第一行：对抗样本；第二行：去噪后的样本)。下标 “T” 表示相应的攻击是目标攻击。

表 2 防御不同对抗攻击生成的对抗样本的对抗准确率。APE-G 和 HGD 使用 PGD_N 生成的对抗样本作为对抗性训练数据

Dataset	Defense	NONE	PGD_T	PGD_N	AA_N	FWA_N	CW_N	DDN_N	STA_N	STA_T
MNIST	APE-G	98.43	96.80	91.24	87.60	66.05	97.66	97.85	83.43	78.60
	HGD	98.64	98.70	98.11	97.57	49.57	98.33	98.46	78.68	63.59
	AIFD	**98.84**	**98.71**	**98.15**	**97.62**	**74.21**	**98.55**	**98.72**	**90.92**	**86.27**
CIFAR-10	APE-G	76.92	60.91	55.62	39.91	57.21	76.82	67.61	52.81	63.54
	HGD	89.59	76.97	60.56	57.66	62.33	86.74	83.98	57.11	68.03
	AIFD	**91.79**	**79.57**	**61.34**	**61.06**	**75.83**	**88.53**	**85.36**	**63.19**	**76.38**

责任编委 崔海楠

参考文献

[1] Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. Explaining and Harnessing Adversarial Examples. ICLR 2015.

[2] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, Rob Fergus. Intriguing Properties of Neural Networks. ICLR 2014.

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR 2018.

[4] Francesco Croce, Matthias Hein. Reliable Avaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. ICML 2020.

[5] Nicholas Carlini, David Wagner. Towards Evaluating the Robustness of Neural Networks. SSP 2017.

[6] Jérôme Rony, Luiz G. Hafemann, Luiz S. Oliveira, Ismail Ben Ayed, Robert Sabourin, Eric Granger. Decoupling Direction and Norm for Efficient Gradient-based L2 Adversarial Attacks and Defenses. CVPR 2019.

[7] Kaiwen Wu, Allen Houze Wang, Yaoliang Yu. Stronger and Faster Wasserstein Adversarial Attacks. ICML 2020.

[8] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, Dawn Song. Spatially Transformed Adversarial Examples. ICLR 2018.

[9] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, no. 11 (1998): 2278-2324.

[10] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, Yongdong Zhang. APE-GAN: Adversarial Perturbation Elimination with GAN. ICASSP 2019.

[11] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba. Learning Deep Features for Discriminative Localization. CVPR 2016.

[12] Alexia Jolicoeur-Martineau. The relativistic discriminator: A Key Element Missing from Standard GAN. arXiv preprint, arXiv:1807.00734, 2018.

[13] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. 2009.

[15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, Brendan Frey. Adversarial Autoencoders. arXiv preprint arXiv:1511.05644, 2015.

[16] Diederik P Kingma, Max Welling. Auto-encoding Variational Bayes. ICLR 2014.

[17] Chuan Guo, Mayank Rana, Moustapha Ciss´e,, Laurens van der Maaten. Countering Adversarial Images Using Input Transformations. ICLR 2018.

[18] Guoqing Jin, Shiwei Shen, Dongming Zhang, Feng Dai, Yongdong Zhang. APE-GAN: Adversarial Perturbation Elimination with GAN. ICASSP 2019.

[19] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, Jun Zhu. Defense Against Adversarial Attacks Using High-level Representation Guided Denoiser. CVPR 2018.

[20] Mortimer Mishkin, Leslie Gail Ungerleider. Contribution of Striate Inputs to the Visuospatial Functions of Parieto-preoccipital Cortex in Monkeys. Behavioural brain research, 6(1):57–77, 1982.

[21] Nancy Kanwisher, Josh McDermott, Marvin M. Chun. The Fusiform Face Area: a Module in Human Extrastriate Cortex Specialized for Face Perception. Journal of neuroscience,17(11):4302–4311, 1997.

周大为

西安电子科技大学在读博士生。主要研究方向是：对抗机器学习及其应用。
Email: dwzhou.xidian@gmail.com

王楠楠

教授，博士生导师，西安电子科技大学，综合业务网理论及关键技术国家重点实验室副主任。近年来从事计算机视觉和统计机器学习方面的研究，在图像跨域重建与可信识别方面进行了深入研究，内容包括跨模态生成、视频理解与分析、底层视觉处理、对抗学习等。
Email: nnwang@xidian.edu.cn

热点追踪

基于多相机系统的全局式三维建模算法

中科院自动化研究所 崔海楠 申抒含

一、摘要

为了充分感知周围三维场景，机器人和自动驾驶汽车等智能体通常会在头顶或者周围安装多相机系统。该系统通过多个相机同步获取 360 度场景数据，然后进行自动三维建模，完成场景三维感知。传统三维建模流程需要利用标定场对多相机系统提前进行离线标定，以解算多个相机之间的相对旋转和相对平移关系；然后，基于该相对位姿关系，再利用增量式从运动恢复结构(SfM)方法对多相机视频数据进行三维建模。然而，多相机系统的离线标定算法通常需要单独的标定场地，以保证多相机之间有足够的共视区域，不仅设计复杂、计算困难而且操作也不方便，一旦更换相机位置或者传感器就需要重新执行离线标定流程；同时，传统的增量式建模方法也无法满足大规模场景三维建模的效率需求。

本文提出基于多相机模型约束的全局式运动平均方法，以解决大规模场景三维感知问题。首先将相机分为参考相机和非参考相机两种类别，从而在计算过程中可以自动地标定多相机之间的相对位姿关系。然后，提出基于多相机模型的旋转平均和平移平均方法，全局式地解决相机的姿态和位置标定问题。该方法不仅操作简单，避免了离线标定的复杂流程，而且极大地提高了场景三维建模的效率和鲁棒性。相关成果已被 AAAI2022 录用为口头报告。

二、多相机模型

在多相机系统采集数据的过程中，由于多个相机是刚性固定的，所以它们之间的相对位姿在运动中是不会发生改变的，即拥有固定的相对旋转和相对平移关系。

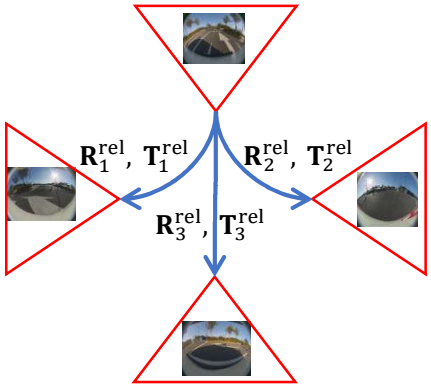

图 1 多相机刚体模型示例

本文将多个相机（假设相机数目为 N）在同一个位置拍摄时的模型定义为一个刚体模型，每个刚体模型中包括 1 个参考相机和 N-1 个非参考相机，那么原多相机系统标定问题就转化为求取非参考相机到参考相机之间的相对旋转$\mathbf{R}_j^{rel}$和相对平移$\mathbf{T}_j^{rel}$。如图 1 所示，多相机系统中包含了 4 个相机，假设前向相机为参考相机，待标定的相对位姿为前向相机与左侧相机、右侧相机和后向相机之间的相对位姿。每个相机的绝对位姿$(\mathbf{R}_j, \mathbf{C}_j)$，可通过多相机模型内部的相对位姿与所在刚体模型的参考相机绝对位姿$(\mathbf{R}_j^{ref}, \mathbf{C}_j^{ref})$计算：

$$\mathbf{R}_j = \mathbf{R}_j^{rel}\mathbf{R}_j^{ref},$$

$$\mathbf{C}_j = \mathbf{R}_j^{ref^T}\mathbf{T}_j^{rel} + \mathbf{C}_j^{ref}。$$

传统 SfM 系统优化参数包括：相机内参和外参，通常采用 5 自由度相机内参模型（包括焦距，两个主点和两个畸变），6 自由度相机外参（相机旋转矩阵和位置）。假设 N 个相机拍摄，每个相机采集 M 幅图像，总参数

表 1 本文方法 MRA 和 MTA 与传统方法的精度对比评测（与真实值对比的误差均值）。其中 RRA、IRA 和 MRA 是全局旋转平均算法，误差评测单位为角度；LUD，BATA 和 MTA 是全局平移平均算法，误差评测单位为米。

KITTI	00	01	02	03	04	05	06	07	08	09	10
RRA[1]	1.7	2.8	88.8	0.7	0.4	1.8	57.8	2.9	0.8	1.4	1.2
IRA[2]	**0.8**	1.6	5.0	36.7	**0.3**	1.3	**0.5**	**0.5**	**0.7**	1.1	**0.7**
MRA	**0.8**	**1.1**	**2.3**	**0.7**	**0.3**	**0.8**	**0.5**	0.6	0.8	**0.6**	0.8
LUD[3]	38.3	41.6	247.0	7.0	66.9	26.4	71.4	13.3	32.2	49.9	40.3
BATA[4]	36.8	29.3	259.9	10.0	87.8	19.7	65.1	8.9	24.8	38.5	26.0
MTA	**1.9**	**23.2**	**10.1**	**1.9**	**13.3**	**1.2**	**1.0**	**2.2**	**3.1**	**3.5**	**1.2**

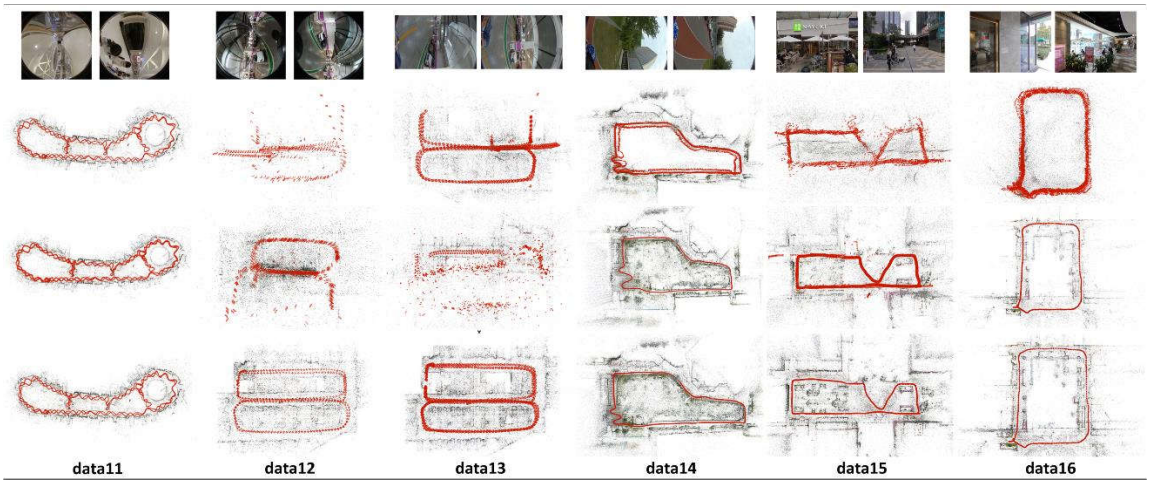

图 2 从上到下依次是 LUD、BATA 和 MTA 建模结果。数据集为 Insta360 多相机采集，红色点表示相机位置。

数目为 5N+6MN。经过参考相机和非参考相机的分类，本文只需要估计参考相机外参和多相机系统内部的相对旋转和相对平移，因此优化参数数目降为 5N+6M+6(N-1)。考虑到大规模场景中拍摄图像数目要远大于相机的数目，那么基于多相机模型的建模系统参数数目约等于传统建模系统参数数目的 1/N。参数的减少不仅可以极大地提高三维建模中捆绑调整优化效率，而且可以提高场景建模的鲁棒性。

三、多相机系统全局运动平均

全局式运动平均流程包括：特征点检测与匹配，摄像机旋转平均，摄像机平移平均，三角化和捆绑调整。基于多相机模型，本文在特征点匹配，摄像机旋转平均和平移平均以及捆绑调整中做了针对性的改变和升级。

对于特征点匹配，由于多相机系统通常都是视频采集数据，所以本文采用的策略是序列式匹配和回环匹配。序列式匹配只对相邻视频帧进行匹配，而回环匹配采用基于图像检索的方式对图像的相似图像进行匹配，实验中采用的是序列式相邻 20 帧图像和检索 50 幅相似图像进行匹配。由于人造场景中可能存在大量的重复纹理，回环匹配中可能存在大量错误匹配。因此在图像匹配完成以后，本文利用多相机模型去衡量检索图像的可靠性：当两幅图像是回环匹配，那么它们所在的多相机刚体也需要有足够多的匹配，即不仅在某一个方向上有匹配，还需要在 360 度的方向有足够的匹配，才可以确定是回环匹配。在特征点匹配完成后，构造场景图，其中图中每个节点表示一幅图像，当两幅图像之间有足够多的匹配点时连接一条边，同时利用本质矩阵分解得到这条边上的相对旋转和相对平移。其中相机相对位姿$(\mathbf{R}_{ij}, \mathbf{t}_{ij})$和绝对位姿$(\mathbf{R}_i, \mathbf{C}_i)$的关系如下：

$$\mathbf{R}_{ij} = \mathbf{R}_j \mathbf{R}_i^{\mathrm{T}}$$

$$\lambda_{ij}\mathbf{t}_{ij} = \mathbf{R}_j(\mathbf{C}_i - \mathbf{C}_j)$$

基于多相机模型的参考相机和非参考相机的分类，原场景图中边转化为四种不同种类的连接边，即参考相机-非参考相机，参考相机-参考相机，非参考相机-参考相机以及非参考相机-非参考相机。每一种连接边上的相对位姿和绝对位姿的关系约束需要重新计算，即公式左侧为本质矩阵分解得到的相对几何关系保持不动，右侧相机的绝对位姿需要利用多相机模型的相对位姿和所在刚体的参考相机位姿表示，从而公式右侧均为要估计的参考相机位姿和多相机模型的相对位姿。根据转化后的场景图中的相对旋转与绝对旋转的关系，将相机旋转矩阵转化为李代数空间 SO(3)进行基于 L1 误差范数的线性求解，然后再利用 L2 范数进行非线性迭代加权求解，得到所有参考相机的绝对姿态和多相机模型的相对旋转矩阵。基于该结果，所有非参考相机的旋转矩阵可以通过第 2 节中介绍的多相机模型计算得到。这里我们将这种方式称为基于多相机模型的全局旋转平均算法 MRA。同理，根据转化后的场景图中的相对平移与绝对位置的关系，我们将等式左右相减，基于 L1 误差范数估计多相机模型的相对平移和参考相机的绝对位置，进而获得非参考相机的绝对位姿，这种方式被称为基于多相机模型的全局平移平均算法 MTA。

本文对多组多相机模型进行测试，包括双目相机(KITTI 数据)，利用 Insta360 采集的双鱼眼和六鱼眼数据。表 1 展示了我们的方法与传统方法对 KITTI 数据的建模精度测试，可以看出我们方法在精度和鲁棒性上较传统方法有了很大提升。图 2 展示了不同方法在其他多相机系统采集数据中的结果，可以看出本文方法鲁棒性更强，更多实验对比结果请参考原文 MMA: Multi-camera Based Global Motion Averaging (aaai.org)。

通过多相机模型理论分析和对大量数据测试，本文方法对于多相机系统内的相机数目和分布均不敏感，且鲁棒性较传统系统有了很大提升。对于机器人和自动驾驶汽车等主流应用，我们的方法可以更加快速鲁棒地进行场景三维感知。

责任编委 王金甲

参考文献

[1] RRA: Chatterjee, A.; and Govindu, V. M. 2017. Robust relative rotation averaging. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 40(4): 958–972

[2] IRA: o, X.; Zhu, L.; Xie, Z.; Liu, H.; and Shen, S. 2021. Incremental Rotation Averaging. International Journal of Computer Vision (IJCV), 129: 1202–1216.

[3] LUD: Ozyesil, O.; and Singer, A. 2015. Robust camera location estimation by convex programming. In CVPR. IEEE

[4] BATA: Zhuang, B.; Cheong, L.-F.; and Lee, G. H. 2018. Baseline desensitizing in translation averaging. In CVPR. IEEE

崔海楠

中科院自动化研究所副研究员。主要研究方向为基于图像的大规模场景三维重建。
Email: hncui@nlpr.ia.ac.cn

申抒含

中科院自动化研究所研究员。主要研究方向为三维重建，机器人三维感知和语义建模。
Email: shshen@nlpr.ia.ac.cn