比较二维与增强现实可视化对微服务系统可理解性的影响：一项受控实验

Amr S. Abdelfattah贝勒大学计算机科学系美国德克萨斯州韦科 amr_elsayed1@baylor.edu

Tomas Cerny贝勒大学计算机科学系美国德克萨斯州韦科 tomas_cerny@baylor.edu

Davide Taibi坦佩雷大学芬兰坦佩雷奥卢大学芬兰奥卢 davide.taibi@oulu.fi

Sira Vegas马德里理工大学西班牙马德里 svegas@fi.upm.es

arXiv:2303.02268v1 [cs.SE] 3 Mar 2023

摘要—基于微服务的系统通常较为复杂，尤其是在规模不断扩大时。从特定视角呈现抽象视图，有助于从业者理解系统。交互式数据可视化的最新进展引出了一个问题：对于微服务的面向服务设计，既有的软件工程模型是否仍是可视化系统设计的最佳方法。我们近期的工作提出了在增强现实中以三维可视化呈现微服务。本文分析这种方法在处理与系统设计质量相关的特定架构问题时，是否能为从业者带来益处。为此，我们开展了一项受控实验，邀请 20 名参与者，考察他们识别服务依赖关系、服务基数和瓶颈的表现。结果表明，在检测服务依赖关系方面，三维可视化使新手能够达到与专家同等的表现，尤其是在大型系统中；而在识别服务基数和瓶颈方面，未发现差异。我们建议工业界和研究人员进一步研究增强现实在微服务架构分析中的应用，尤其是利用它帮助新开发人员在微服务项目中开展入职培训。

索引词—微服务、可视化、增强现实、服务依赖关系图、受控实验

I. 引言

云原生设计与涉及众多相互作用以解决企业问题的自包含微服务的复杂系统相关。其主要优势之一是去中心化，这使独立团队能够开发和维护自包含微服务，从而提供更强的自主性并支持独立部署。同时，它也带来了微服务面临的一项重大挑战：随着去中心化的发展，我们失去了有助于实现可观测性和全面理解系统的以系统为中心的视角。这种视角对于推动高效的系统演进非常有用。

面向系统的视角中，各种架构视图可以大幅抽象底层系统的复杂性。在推理微服务依赖关系或其他视角时，它们能够减少不必要的细节（即领域、技术和运维）。我们可以分析现有系统并提取这些视图，从而更好地分析其架构在演进过程中的变化 [1]。提取过程可以手动完成，也可以借助动态或静态分析。鉴于微服务源自面向服务的架构，许多用于监控云原生系统的现有工具都广泛采用系统的服务视图，因此我们聚焦于这一视图。

然而，现有工具采用一体化的静态服务视图可视化，缺乏交互性 [2], [3]。此外，我们必须考虑到，就不同微服务的数量而言，系统规模可能会超出既有可视化模型的设计范围（即单体系统）。问题在于，增强现实（AR）及其交互性转变，是否能够让从业者在使用既有可视化方式解决应用于微服务的服务视图中的常见质量推理任务时，表现得比使用传统可视化方式更好。

本文详细阐述了服务视图可视化。我们近期提出的服务视图可视化方法采用在增强现实（AR） [4]中呈现的三维节点-边模型。提出这一可视化模型的依据是，三维空间能够更好地容纳复杂图，而且与使用成熟的传统二维可视化模型相比，终端用户在有限空间内可能更容易阅读该模型。

我们的工作旨在比较在分析以二维或增强现实呈现的服务调用图时，人们对基于微服务的系统的感知易理解性。为此，我们在 20 名开发人员中开展了一项受控实验，比较他们在检测两种不同规模的系统中的服务依赖关系、服务基数和瓶颈时的表现。

此外，我们还旨在了解从业者会如何接受新的增强现实模型，以及使用增强现实来更好地理解技术转变的影响时会面临哪些挑战。

结果表明，增强现实可视化使新手开发人员能够像有经验的开发人员一样理解系统依赖关系。此外，与二维工具相比，该方法在可视化大型系统方面效果显著，而二维工具的效果则会相应下降。

这项工作的研究结果可为研究人员提供帮助，使其进一步开发和评估面向微服务的增强现实可视化；也可为工业界提供参考，以扩展现有工具，将增强现实集成到工具中。例如，Jaeger (https://www.jaegertracing.io) 或 Kiali (https://kiali.io) 等跟踪工具，或许可以轻松地在其二维服务视图上集成增强现实。

本文的组织结构如下。第二节介绍相关工作和微服务可视化。第三节陈述研究问题以及用于回答这些问题的研究变量。第四节详细介绍本研究及其实验设计。第五节呈现研究结果，第六节对结果进行讨论。第七节说明本研究的有效性威胁，第八节给出结论。

II. 相关工作

软件架构重构（SAR） [5] 旨在分析现有系统，以获取该系统已实现的架构。架构是系统蓝图，也是基于组件的系统 [5] 开发与设计的核心。根据 O'Brien 等人 [5]，的观点，该过程的结果可用于评估实际架构是否符合文档化架构；也可用于文档不完善的系统，或用于分析和理解现有系统的架构，以便修改架构来满足新需求并消除现有软件缺陷，这体现了系统的可演化性 [6]。此外，该过程还用于系统验证和权衡分析 [1]。

对于微服务中的软件架构重建 [1]，[7]，[8]，，我们会从系统代码库、源代码，甚至容器描述文件、构建文件或许多配置文件中查找相关系统信息。在手动重建时，也可以使用现有文档 [1]，[7]。我们还可以利用系统运行时和监控来确定微服务之间的依赖关系 [9]，[10]。微服务依赖图已用于测试微服务 [11]，，这表明动态分析可以成为一种可行途径，从特定视角对系统进行整体分析。

软件架构对不同的专家而言可能意味着不同的内容，因此我们通常会考虑不同的视点。针对微服务架构，已有若干此类视图 [1]，[7] ，例如领域视图，通过领域模型和微服务限界上下文处理领域专家所关注的问题；技术视图，用于识别微服务中所采用的技术；服务视图，用于构建描述微服务、接口和端点的服务模型；以及运维视图，关注服务部署的拓扑结构以及与服务编排、服务发现、监控和其他运维问题相关的基础设施。

尽管软件架构重建有多种方法，但我们并不总是需要全面的架构概览。聚焦于某个选定视角可能就足够了。例如，分析微服务依赖图以对其进行测试可能就足够了 [11]，，为此我们可以采用简化流程。同样，现有的跟踪和监控工具通常只能提供基础的服务概览（即服务调用图）。接下来，自然是以针对性的方式向从业者展示整个系统的可视化结果。这样，他们就能更好地理解系统，并完成使用这些工具所要执行的任务。因此，本实验旨在衡量不同可视化方法对从业者理解系统的影响。

A. 微服务可视化

微服务语境下出现的一个问题是，如何将系统视点可视化。UML 和系统建模语言（SysML）等传统可视化模型都是衍生模型，但微服务增加了需要考虑的各种方面，其中最主要的是系统规模。问题在于，在微服务架构出现之前所使用的现有可视化技术，是否适用于当前的挑战 [3]，[12]。尽管我们可以毫不费力地使用 UML类图来表示微服务语境下的领域，但如何呈现服务视图仍然令人关注。The OpenGroup Architecture Framework (TOGAF) [13] 和 ArchiMate 企业架构建模语言 [14]早在将近二十年前微服务出现之前就已存在。

微服务研究人员观察到，就微服务而言，系统的整体视图应包含服务API及其交互 [15]。软件架构重建的服务视图描述了微服务的API及其之间的服务间调用。此外，服务视图可能有助于 DevOps 和开发人员了解依赖关系及相互连接，从而评估其演进过程中可能产生的连锁影响 [6]。

数据可视化引入了可供利用的替代途径；其中包括用于语义网的三维节点边图、采用各种隐喻（即岛屿、城市）的系统和分布式系统中的 [16]，虚拟现实，以及增强现实 [19]。微服务架构，尤其是服务视图和不断扩大的系统规模，似乎正朝着更高效的空间利用方向发展，而这些方法可以提供这种空间利用。

一种应用于微服务的方法是 Microvi- sion [4]，，它将三维节点边图与增强现实结合起来。Microvision 考虑每个服务的端点，并连接通过这些端点进行交互的服务。这些信息可以通过跟踪工具获取，也可以借助 Open API（ https://www.openapis.org，正式名称为 Swagger https://swagger.io），并结合通过 REST模板或类似方式识别远程方法调用来获取。动态分析在识别实际远程调用方面可能更加精确。这是因为静态

分析方法在匹配调用与端点签名及 HTTP 方法时存在局限（还可能需要结合目标服务名称的配置信息）。然而，动态分析需要全面的流量，才能识别所有追踪组合。如果追踪记录不完整，我们可能无法发现完整的服务调用图。服务视图中的微服务视角会呈现一个服务调用图，将系统拆分为各个微服务及其具体端点，并展示微服务之间的通信方式。

Microvision 运行于增强现实环境中，这改变了我们使用既有可视化模型时熟悉的服务视图导航方式。我们不再使用个人计算机、键盘和鼠标，而是需要使用配备摄像头的设备，在渲染出的系统图中进行导航。此外，这种表示方式还增加了更多交互功能：我们可以选择某个微服务，以突出显示其连接关系，并分层详细展示微服务端点及其属性。

挑战在于，人们无法判断这种方法是否能够帮助从业者更好或更有效地应对识别架构属性、质量属性和服务依赖关系时的常见任务。此外，这种技术转变还可能带来新的实际障碍。

III. 研究问题与研究变量

本研究评估了两种基于微服务的系统可视化方法：二维可视化和基于增强现实（AR）的三维可视化。因此，本研究考察了增强现实方法对基于微服务的系统可理解性的影响。此外，本研究将系统规模和从业者经验水平作为使用这些工具时的主要影响因素。系统规模 体现了微服务数量对系统可理解性的影响。因此，本研究使用小型和大型系统变体对这些工具进行考察。另一方面，从业者经验水平 衡量不同开发经验水平对系统可理解性的影响。本研究根据参与者在微服务开发方面的经验水平，将其分为新手组和有经验的组。总之，本实验的目标表述如下：

> 评估两种可视化方法，以衡量基于微服务的系统的可理解性，并考察系统规模和从业者的经验水平。

本文提出并探讨以下研究问题（RQ），以实现这一目标：

RQ1. 与增强现实相比，二维可视化是否更适用于理解基于微服务的系统？

本研究旨在引入一些特定任务，用于评估不同方面的可理解性。系统。因此，本研究通过可视化工具，从以下角度评估系统的可理解性：

- **依赖关系：** 衡量使用给定可视化工具识别系统中微服务之间依赖关系的能力。
- **基数：** 衡量识别微服务之间依赖程度的能力。
- **瓶颈：** 考察参与者在多大程度上能够检测出整个系统中依赖程度最高的微服务。

研究问题***2***。参与者对于使用增强现实来理解基于微服务的系统有何口头表达的看法？

该问题涉及参与者执行任务时对工具使用的反馈。

评估过程强调了回答该问题的以下标准：

- **易用性：** 该标准衡量参与者使用工具所需付出的努力，以及查找所需任务答案的难易程度。
- **完整性：** 该标准衡量查找理解系统和回答任务所需全部信息的能力。
- **推荐：** 该标准衡量参与者向他人推荐该工具用于日常工作环境的程度。

RQ3. 参与者认为基于增强现实的技术面临哪些挑战？

这个问题涉及参与者对所研究工具的看法。评估与研究相关的挑战需要分析各种测量标准。因此，研究收集了关于多个标准的文字意见，以帮助识别挑战以及可用于克服这些挑战的建议。这些标准通过比较工具中最棘手的问题与最容易处理的问题来进行概括，也就是说，衡量参与者为绕过任务所付出的努力。

以下各节将讨论研究详情，并阐述这些研究问题的答案。

IV. 实验设计

本研究衡量了两种可视化方法对基于微服务的系统可理解性的影响。以下各小节将介绍研究设置流程，直至得出数据分析结论。研究方案[1] 包括材料和脚本，均由作者准备，并经贝勒大学机构审查委员会（IRB）审查（#1845572）。

[1]研究方案：https://zenodo.org/record/7693694#.ZAEi_S2B3RY。

表 I 参与者分布

参与者经验		系统 Size	会话 1	会话 2	组
P1	新手	大型	2D	–	G1
P2	有经验的	大型	2D	–	G1
P3	新手	大型	2D	AR	G1
P4	新手	大型	2D	AR	G1
P5	新手	大型	2D	AR	G1
P6	有经验的	小型	2D	–	G1
P7	经验丰富的小型		2D	–	G1
P8	经验丰富的小型		2D	–	G1
P9	经验丰富，大型		2D	AR	G1
P10	有经验的	小型	2D	AR	G1
P11	新手	小型	2D	AR	G1
P12	有经验的	大型	2D	AR	G1
P13	经验丰富，大型		AR	2D	G2
P14	新手	大型	AR	2D	G2
P15	有经验的	小型	AR	2D	G2
P16	新手	小型	2D	AR	G1
P17	经验丰富的小型		AR	2D	G2
P18	新手	小型	–	2D	G2
P19	新手大型		AR	2D	G2
P20	经验丰富的小型		AR	2D	G2

A. 可视化工具

本研究引入了两种可视化工具：二维和增强现实。二维工具是一种成熟的可视化方式，与商业和开源工具中使用的方式类似。它使用矩形框和箭头来表示微服务依赖图中的服务及其调用，如图1所示。另一方面，增强现实工具通过立方体表示服务，并使用带有调用信息弹出窗口的线条连接器，来渲染服务依赖图的三维模型，如图2所示。尽管两种可视化方式提供的信息相同，但它们提供了不同的显示和交互选项，以满足不同需求。二维方法适合静态且可能需要打印的视图，而增强现实则支持更广泛的交互，并可视化多个同一视图中的信息片段。

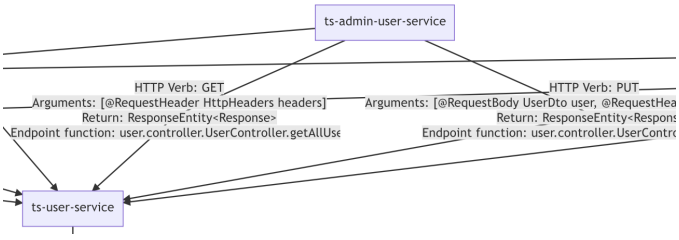

图1. 二维工具（完整版本可在 protocoldata 中获取。）

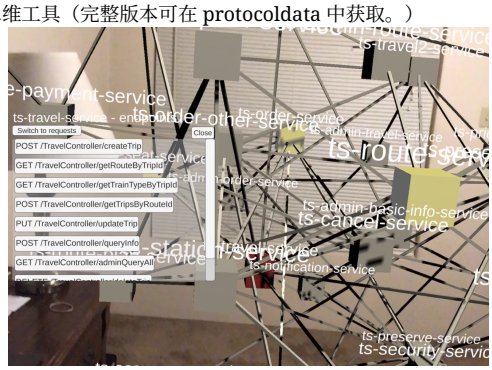

图2. ARTool（完整版本可在协议数据中获取。）

这两种可视化在呈现相同信息时各有特点。二维工具是一款基于 Web 的应用程序，可通过台式电脑和笔记本电脑访问。另一方面，增强现实可视化需要使用屏幕空间有限、交互方式不同的设备。它专为安装在手机或平板设备上而设计，并通过移动设备摄像头在增强现实空间中呈现三维表示。增强现实应用程序通过 APK 文件分发到安卓设备（至少需要 Android 8），并通过 IPA 文件分发到 iOS 平台（至少需要 iOS 11）。

两种工具均使用 TrainTicket（V 0.0.1） [20] 测试基准作为被测系统。该系统基于微服务架构，包含 41 个 Java Spring 微服务。使用 Prophet 工具 [21] 提取相应的服务依赖图，并考虑了两种系统规模：大型系统包含全部 41 个微服务，小型版本则包含 16 个微服务。小型版本通过调整图来生成，同时保留微服务之间的连接关系，从而复现微服务系统架构。这使研究能够考察不同系统规模对工具的影响。

B. 参与者

研究对象包括来自美国和欧洲各机构的20名参与者。研究开始前，我们通过问卷收集了背景信息，例如软件工程和微服务经验、微服务项目数量、项目规模等。问卷通过电子邮件发送给26名英语候选人，其中23人有意参与，最终有20人完成了研究。参与者的筛选标准是具有一至五年的微服务相关开发经验，因为这一经验与本研究最为相关。尽管招募从业者颇具挑战性，本研究仍选择优先为从业者获取有价值的成果，而非招募更多新手。

这些参与者根据其微服务开发经验年限分为两组。经验少于两年的参与者被归类为新手；否则，则视为有经验者。第一组（新手组）由九名新手开发人员组成，平均经验为一年；第二组（有经验组）有十一名有经验的开发人员，平均经验为四年，表 I 展示了参与者的分组情况。在研究正式开展之前，另有两名参与者自愿参加了试点研究，但他们的数据未纳入正式研究。开展试点研究是为了验证实验的有效性。试点研究使团队能够据此修改和改进培训材料、时间安排、任务及问题。

表 II 全尺寸系统（41 个服务）的研究任务

二维任务	增强现实任务
依赖关系	
1）列出依赖于“ts-order-service”的服务。 **答案：** ts-cancel-service, ts-admin-order-service, ts-seat-service, ts-preserve-服务, ts-execute-service, ts-inside-payment-service, ts-travel-服务，ts-security-service。	1) 列出依赖于服务“ts-travel-service”的服务 **答案：**ts-admin-travel-service、ts-route-plan-service、ts-preserve-service、ts-travel-plan-service、ts-seat-service、ts-food-service。
2) 列出服务“ts-travel-service”所依赖的服务。 **答案：** ts-ticketinfo-service ，ts-seat-service ，ts-train-service ，ts-route-服务 ，ts-order-service。	2）列出依赖于服务“ts-order-service”的服务。 **答案：**ts-station-service。
基数	
3）有多少个服务依赖于服务“ts-order-service”? **答案：**8 个服务。	3）有多少个服务依赖于服务“ts-travel-service”? **答案：**6 个服务。
4) 服务「ts-travel-service」依赖多少个服务 on? **答案：**5 个服务。	4) 服务「ts-order-service」依赖多少个服务 on? **答案：**1 个服务。
5) 请列出两个至少被 4 个微服务依赖的微服务（取决于它）。 **答案：**ts-seat-service ，ts-travel-service ，ts-user-service ，ts-route-service ，ts-station-service、ts-order-service。	5）列出一个至少有 5 个依赖它的微服务的微服务（依赖于它），并说明这些微服务的确切数量。 **答案：**ts-travel2-service ，ts-admin-basic-info-service ，ts-route-service ，ts-preserve-service。
瓶颈	
6) 依赖关系最多的微服务是什么? **答案：**ts-preserve-service（12 个依赖关系）。	6) 依赖关系最多的微服务是什么? **答案：**ts-preserve-service（12 个依赖关系）。

表 III 方面 反馈 五点李克特量表排名（1 为最差，5 为最好）

易用性
1) 可用性（易于使用）
2) 可理解性（易于理解系统）
3) 获取信息所需时间
完整性
4) 包含所有所需信息
推荐
5) 您认为在日常工作中采用这种方法的可能性有多大?

表 IV
反馈段落问题

请用文字回答这些问题
1) 使用这种方法，您可以轻松获取哪些信息?
2) 你觉得通过这种方法难以获取哪些信息?
3) 你遇到的最大痛点是什么?
4) 你对改进这种方法有何建议?

C. 材料

本研究涵盖了多种材料，首先是通过电子邮件发送的参与意向问卷，用于征询候选人是否有兴趣参与研究。在线表单旨在收集参与者的相关信息，包括他们的经验，以及偏好的研究日期和时间。此外，每种可视化方法都包含三份根据相应内容定制的文档。这些文档用于说明研究的执行流程，如下所述：

• 培训文档：使用了一个 PDF 文件，详细介绍了相应的工具及其所需的安装流程。文件还展示了预期任务的示例，并附有一段简短视频来演示该工具的使用。

•任务表单：包含六个问题的在线表单（任务）。这些任务在不同工具之间有所差异，以避免答案记忆对结果的影响。任务难度相同，且聚焦于各种微服务的相同特征，因此会得出不同答案。例如，不同版本的任务使用相同的问题，但涉及系统中的不同微服务。这些任务分为三类（见 表 II）：依赖关系、基数和瓶颈。如上文研究问题 1所述，这些类别有助于描述和展示系统的可理解性。

• 反馈表：在线表单要求参与者在完成任务阶段后对该方法进行评估。表单包含两类问题：如表 III 所示的五点李克特量表问题，以及表 IV 所列的段落式问题，以获取文字描述性评价。这些反馈为第III节中所述的研究问题2和研究问题3提供答案依据。

这三份文档和表单被整合到一份说明文档中。该 PDF 文件简要介绍了每个系统及其包含的微服务数量，并包含对培训文档、任务表单和反馈表单的引用；同时指导参与者如何执行研究，以及完成后如何提交结果。每份方法说明文档都会根据预定时段通过电子邮件发送给相应的参与者。

D. 设计流程

实验采用 2×2交叉设计，每个参与者组都使用两种可视化工具。每位参与者会在其为研究预约的时间段收到一封电子邮件。该邮件包含两份有序脚本，参与者必须按相同顺序执行；每份脚本都包含第 IV-C 节所述特定方法的完整流程。执行顺序

表V 可理解性变量的显著性统计

变量	依赖关系 (P值)	基数 (P值)	瓶颈 (P值)
截距	<0.001	<0.001	<0.001
经验水平	0.986	0.371	0.218
系统规模	**0.041**	**0.002**	0.089
组	0.446	0.562	0.618
会话	0.067	0.711	0.640
Tool	0.515	0.158	0.556
工具 * 经验水平	**0.035**	0.606	0.747
工具 * 系统规模	**<0.001**	0.132	0.460

在不同参与者之间交叉安排；组列表 I 指明参与者先使用二维还是增强现实进行实验。此外，每个时间段都有两位作者值守，为参与者提供支持并解答他们在实验过程中提出的任何疑问。

实验时长为一小时，分为十分钟的培训、30分钟的实验执行与问题作答（任务），以及十分钟的反馈填写与提交。每位参与者都必须使用每种工具进行一次实验环节。所有参与者接受了相同的必要知识与培训，内容涉及实验、目标、接口、研究时间和方法。

实验从十分钟的培训开始，培训通过培训文档进行。培训确保参与者做好准备，并在开始执行任务前熟悉应用程序及任务的性质。培训还指导参与者如何在自己的设备和浏览器上安装并运行应用程序。随后，培训展示了一个应用程序使用用例，并要求参与者回答一个与任务中问题类似的问题。之后，参与者用30分钟完成任务，回答表 II中列出的任务相关问题。任务答案用于衡量系统在依赖关系、基数和瓶颈这三个方面的解读准确性。因此，根据从测试平台系统中提取的实际答案，评估每位参与者的答题正确率。我们分析这些数据，以展示研究变量对系统可理解性的影响。

最后，提交任务后，反馈环节会被突出显示。该环节由两部分组成：使用相关标准和开放式意见问题。第一部分的问题见表 III，采用五点李克特量表进行等级评分（1 为最差，5 为最好）。这些数据将在后续章节中进行统计分析。第二部分要求参与者写下对表 IV中问题的看法。我们对这些反馈进行分析，以提取常见观点和建议。

本研究旨在根据系统规模和参与者经验进行均衡分组，纳入十个小型系统和十个大型系统，并为每个系统分配特定数量的新手和有经验的参与者。为避免其中一种工具产生比另一种更大的影响，我们用抽象名称（如工具 1 和工具 2）指代这两种工具，以尽量减少名称对参与者可能产生的影响。此外，我们随机安排了两种工具对应脚本的执行顺序，使每组中的部分参与者以与其他人相反的顺序完成研究。每位参与者使用工具的顺序见实验环节 *1* 和 实验环节 *2* 列，详见表 I，其中 组 列将先使用二维的参与者标记为 G1，将先使用增强现实的参与者标记为 G2。

由于六名参与者的移动设备存在兼容性问题，他们难以使用增强现实工具。因此，四名小型系统参与者和两名大型系统参与者未能获得增强现实工具的结果。于是，对参与者的分组稍作调整，如表 I所示。具体而言，两名新手参与者从大型系统组调至小型系统组，一名有经验的参与者从小型系统组调至大型系统组，这些调整在参与者列中以粗体标出。此外，14 名参与者完成了两个实验环节，因此 G1 进行了五次大型系统实验和三次小型系统实验，而 G2 进行了三次大型系统实验和三次小型系统实验。

总之，对实验流程的评估表明，所有参与者都能够成功完成二维实验。不过，六名参与者因其移动设备不符合操作系统要求（最低要求为 Android 8 或 iOS 11），无法安装增强现实工具。尽管如此，他们的结果仍纳入二维工具的结果中，但在增强现实工具的数据中被视为缺失数据。收到的参与者电子邮件咨询均与增强现实工具的安装问题有关，除此之外，研究进展顺利，没有其他问题。

E. 分析方法

首先，我们将报告描述性统计结果，随后进行假设检验。RQ1 和研究问题2 使用不同的检验方法。

1) RQ1： 按照 Vegas等人 [22]，的做法，我们在本实验中使用线性混合效应模型分析 RQ1。该模型用于检验工具（二维/增强现实）、参与者的经验水平（新手/专家）和系统规模（小型/大型）的影响，以及经验水平与工具、系统规模与工具之间的二阶交互作用对因变量——依赖关系、基数和瓶颈——的影响，同时控制所选交叉设计（会话和组）可能产生的影响。

我们使用 SPSS v.27 MIXED过程分析数据。我们采用两种数据表示方式拟合数据：一种将参与者视为随机因子，另一种将工具视为被试内因子。对于后一种情况，我们拟合了五种不同的协方差结构（单位结构、对角结构、一阶自回归结构、复合对称结构和非结构化结构）。因此，每个因变量都需要拟合六个不同的模型。

我们使用赤池信息准则（AIC）衡量模型拟合度。AIC 值本身并无大小之分，但可以在不同模型之间进行比较。值越小，拟合效果越好，因此我们将选择 AIC 最低的模型。

MIXED 程序要求残差服从正态分布。我们通过正态概率（Q-Q）图对此进行了检查。当数据不满足这一标准时，需要对因变量进行变换。

如果能够确保数据服从正态分布，则效应量还以 Cohen's d 统计量的形式报告 [23]；否则，我们报告非参数效应量 Cliff's delta [24]。

2) 研究问题2：我们使用相关样本的非参数 Wilcoxon 符号秩检验来分析研究问题2。该检验用于检验工具（二维/增强现实）对以下因变量的影响假设：易用性、易理解性、获取信息所需的时间，以及是否包含所有必要信息。

3) RQ3：我们对针对 表 IV 中问题提交的简短回答进行了定性分析。我们阅读了收到的关于二维和增强现实实验的所有反馈。为确保客观性，我们每个人都根据对各个回答的个人理解对反馈进行了分类。随后，我们讨论并合并了相似的类别。之后，我们总结并报告了共同的反馈。

V. 数据分析结果

我们分析了从[1] 名参与者处收集的数据集。针对每项数据执行相应的分析方法，以回答各项 RQ。

A. RQ1. 二维可视化是否比增强现实更适用于理解基于微服务的系统?

我们检验了六个模型，以评估它们对数据的拟合程度。这些模型根据参与者的建模方式进行配置：其中一个模型未指定重复测量，随后将参与者指定为随机变量。另一方面，其余五个模型均指定了重复测量，但检验了不同的协方差类型，例如比例单位矩阵、复合对称、对角、非结构化和自回归。

我们基于赤池信息准则和残差的正态性（观察 Q-Q 图）比较了这些模型的拟合程度。比较结果显示，将参与者配置为随机变量的模型被选中；该模型显示残差呈正态分布（因此不需要进行数据变换），并且在所有模型中具有最佳的赤池信息准则值。该模型应用于代表本研究可理解性属性的因变量，具体包括：依赖关系、基数和瓶颈。

通过 表V 中列出的 P值，可以总结自变量对可理解性的显著性。因变量受到工具与各系统规模及经验水平之间交互作用的影响。尽管系统规模变量显示出显著性，但由于其与工具的交互作用具有更高的优先级，因此我们不考虑其影响。为衡量这种影响，我们绘制变量交互作用的估计边际均值，并计算其效应量。首先，系统规模与工具之间的交互作用凸显了ARtool的可扩展性。对于大型系统，增强现实工具是可视化的最佳选择，且具有较大的效应量（$\delta = 0.96$）；然而，如 图3 所示，两种工具对小型系统都没有显著影响。对于经验水平的显著性，分析结果同样表明，二维工具需要有经验的参与者才能检测系统的依赖关系。不过，这属于较小的效应量（$\delta = 0.19$），如 图4 所示。

基数变量表明，系统规模是唯一的影响因素，而增强现实工具不会影响基数相关问题的回答准确性。不过，分析显示，小型系统有助于参与者更准确地回答基数问题。此外，增强现实不会影响瓶颈结果，因此，相关任务可以选择二维工具或增强现实工具，而不会影响结果。

B. 研究问题*2*：参与者对于使用增强现实来理解基于微服务的系统有何言语表达的感知?

对两种工具的易用性评分进行了比较。平均而言，二维工具的表现（中位数 = 2）不如增强现实工具（中位数 = 3），效应较小（$\delta = 0.19$）。Wilcoxon符号秩检验表明，这一差异在统计上并不显著（T=30，Z=0.263，p = 0.793）。

对两种工具的易理解性评分进行了比较。平均而言，二维工具的表现较差（Mdn =3），而增强现实工具的表现较好（Mdn = 4），且呈现出较大的效应（$\delta = 0.46$），在大型系统中尤为明显。Wilcoxon符号秩检验表明，这一差异具有统计学显著性（T=66, Z=2.164, p = 0.030）。

对两种工具的获取信息所需时间评分进行了比较。平均而言，二维工具的表现较差（Mdn =2.5），而增强现实工具的表现较好（Mdn = 4），且呈现出较大的效应（$\delta = 0.38$）。Wilcoxon符号秩检验表明，这一差异在统计学上不显著（T=67, Z=1.548, p = 0.122）。

对两种工具的所需信息完备性评分进行了比较。平均而言，二维工具的表现较好（Mdn = 4），而增强现实工具的表现较差（Mdn = 3.5），且呈现出较小的效应（$\delta = 0.28$）。Wilcoxon符号秩检验表明，这一差异在统计学上不显著（T =16, Z=-1.218, p = 0.223）。

对两种工具的推荐评分进行了比较。平均而言，二维工具（Mdn = 2）的表现不如增强现实工具（Mdn = 4），且效应量较大（$\delta = 0.57$）。Wilcoxon符号秩检验表明，这一差异具有统计学显著性（T=69，Z=2.435，p = 0.015）。

此外，我们对这些响应变量进行了描述性统计，结果汇总于图5。增强现实和

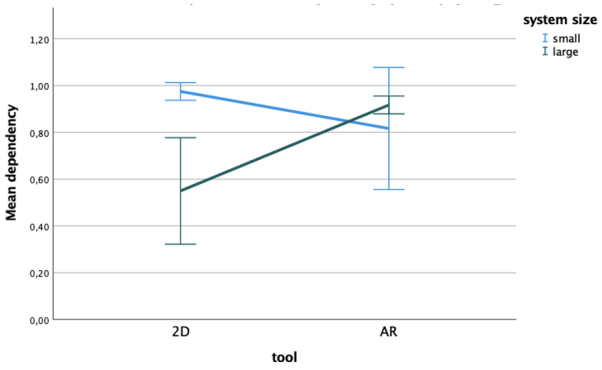

图 3. 系统规模对依赖关系的影响（δ = 0.96）

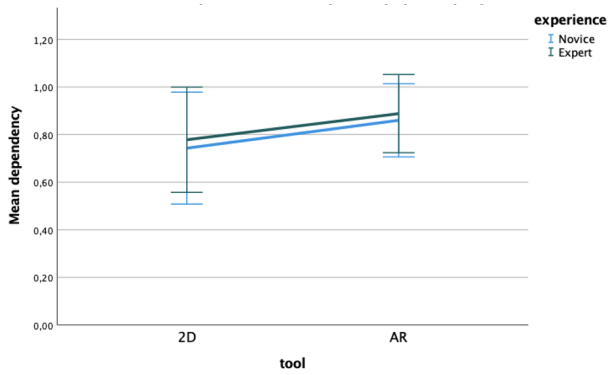

图 4. 经验水平对依赖关系的影响（δ = 0.19）

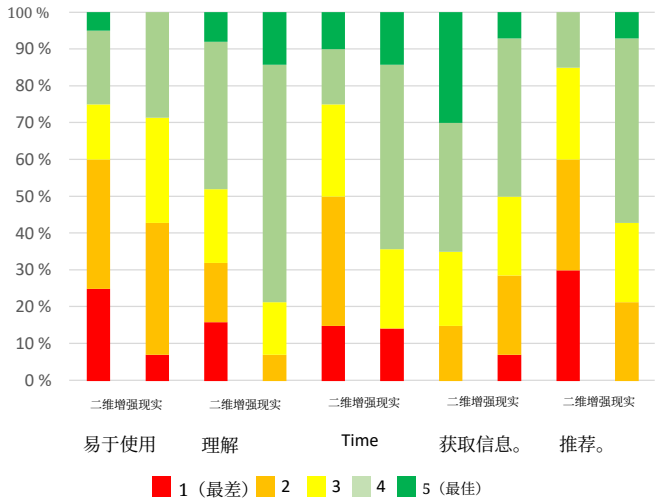

图 5. 关于增强现实和二维的口头看法是什么（研究问题2）

二维工具对易于使用这一响应的影响并不显著（图5 易于使用）。如图5 （获取信息）所示，在“包含所需的全部信息”这一变量上，二维工具的排名优于增强现实工具。然而，在易于理解（图5 理解）以及获取信息所需时间更短（图5时间）这两方面，增强现实工具均优于二维工具。此外，如图5 （推荐）所示，参与者更倾向于推荐增强现实工具，而非二维工具。

C. RQ3. 参与者认为基于增强现实的技术面临哪些挑战？

分析参与者对表 IV中所列问题的评价和回答；他们的回复总结于表 VI。对于二维工具，参与者指出，为了追踪服务之间交叉的连线，他们需要在图中频繁滚动，这令人十分不便。他们建议按需显示信息，例如在用户点击特定节点或边时再显示。对于增强现实工具，参与者表示，他们需要更多控制功能，以便放大视图并重新排列服务节点。此外，两种方法都普遍被建议改进颜色的使用，例如用颜色区分不同的请求类型或依赖方向。参与者表示，以下因素对他们的影响很大：

相较于增强现实工具中较为简单的依赖线，他们认为二维工具中的依赖线很难追踪。他们还表示，增强现实工具通过展示系统的整体概况以及服务之间清晰的依赖关系，增进了他们对系统的理解。

VI. 讨论

这项工作的结果使我们能够了解增强现实在理解微服务架构方面的作用。

在评估增强现实是否适用于检测服务依赖关系时，有经验的开发人员使用两种工具的表现相近。然而，增强现实使新手开发人员在处理小型和大型系统时，都能达到与有经验的开发人员相当的理解水平。因此，相较于增强现实，使用二维识别系统依赖关系需要更高的专业水平。新手开发人员表现有所提升的原因之一，尤其是在大型系统中，是这类系统通常更难理解，而增强现实工具使他们能够以三维方式旋转系统，从而缩放视图或更好地聚焦于系统的特定部分；相比之下，二维系统可能会让新手感到复杂，这或许是因为大量信息被压缩到有限空间中。此外，分析表明，系统规模会显著影响使用所研究工具检测依赖关系的效果。因此，对于小型系统，二维工具对这一特征的可视化比增强现实工具更清晰。然而，在以可扩展的方式可视化大型系统方面，增强现实表现更佳。

出乎意料的是，在分析服务基数时，二维与增强现实之间并未出现差异。无论专业经验如何，参与者都能在两种可视化中准确识别基数。不过，正如预期，基数与系统规模相关。在规模更大的系统中，识别基数更加复杂，所需时间也更长。

同样出乎意料的是，瓶颈识别不受所采用的可视化方法、系统规模或参与者经验的影响。

在考虑参与者对二维和增强现实可视化的看法（研究问题2）时，ARtool 使得

表 VI 段落问题反馈分析（RQ3）

二维工具	增强现实工具
你遇到的最大痛点是什么?	
1. 在示意图中来回滚动。 2. 有许多交叉的通信线。	1. 图周围有很多移动。 2. 增强现实引擎会突然移动图。 3. 服务的位置是固定的。
你对改进这种方法有什么建议?	
1. 使用颜色突出显示依赖关系。 2. 按需显示信息。 3. 提供搜索和筛选功能。 4. 使用对话框显示依赖服务和被依赖服务。	1. 提供缩放和移动节点的控制功能。 2. 用不同的颜色表示另一个方向的依赖关系。 3. 允许自定义颜色。
使用这种方法可以轻松获取哪些信息?	
1. 端点信息。 2. 服务依赖关系基数。	1. 系统概览图。 2. 每个服务支持的端点。 3. 服务之间的依赖链接。
通过这种方法，您觉得哪些信息难以获取?	
1. 服务依赖关系。 2. 依赖关系的方向。	1. 调用端点的详细信息。2. 依赖于特定服务的服务（另一种依赖方向）。

开发人员能够更好地理解所分析的基于微服务的系统。这可能是因为增强现实工具的三维可视化支持在更高渲染空间中呈现不同视点。此外，推荐结果证实了参与者更偏好增强现实工具，因为该工具采用了更清晰的依赖关系表示。

易用性、获取信息所需的时间，以及参与者分析系统所需的信息，均未得出具有统计显著性的结果。不过，即使结果不显著，在大多数情况下，通过增强现实获取信息所需的时间远少于使用二维工具；而在分析基于微服务的系统时，二维工具提供的信息比增强现实更多。这可能是因为增强现实可视化需要额外步骤才能找到所需信息。

总之，我们建议新手开发人员考虑使用增强现实可视化来呈现大型系统中的依赖关系。不过，服务调用图的二维可视化（例如 Jaeger 或 Kiali 提供的可视化）仍可能有助于识别服务基数和系统瓶颈。

因此，考虑到微服务系统软件开发领域缺乏有经验的劳动力，我们强烈建议企业和研究人员进一步开发并验证用于可视化基于微服务的系统的增强现实方法。这也有助于微服务项目中新加入开发人员的入职过程。另一方面，尽管参与者对增强现实工具的推荐程度高于其对照工具（图 5- 推荐）。但也必须考虑到，面对一种非传统方法时，可能会产生初见效应（第一印象）。需要强调的是，增强现实要取得成功，就必须将反馈中提及的多项功能和普遍期望纳入该方法。例如，用户需要快速的搜索和筛选机制。虽然增强现实前景看好，但它也给实际应用带来了限制，例如要求开发人员使用配备摄像头的便携设备。此外，离开工作环境中的工具可能会让人感到有些不适。因此，考虑将可视化工具与开发工作区集成，或许能提升这些工具的可用性。

VII. 效度威胁

我们根据 [25] 提出的分类，讨论了本研究的效度威胁。

1) 构念效度： 本研究遵循对照实验的实践规范和指南来构建、分析并报告研究及其结果。本研究的核心涉及我们的原型工具，这些工具可能会影响结果。不过，两种原型工具的开发都以提升可用性为目标，且开发者秉持良好意图。原型工具的作者与设计用户研究的人员彼此独立，以避免偏见。双方没有就研究设计和实施进行任何交流。此外，这些工具呈现了使用我们的 SAR 流程生成的结果；尽管如此，这款与 SAR 相关的软件已公开发表，并通过多个项目得到验证。

我们使用小型但贴近实际的任务收集结果，这些任务是从业者可能会遇到的，并根据对微服务领域有经验的开发人员的访谈选定。

2）内部效度： 实验内部效度可能受到疲劳效应、携带效应和练习效应的威胁。由于任务执行时间相对较短，因此疲劳效应被认为可以忽略不计。研究以远程方式进行，在安排的实验时段内，两位作者均可通过电话或电子邮件联系；参与者则选择最方便的时间参与研究。由于实验本身的性质，即受试者依次采用二维和增强现实，因此携带效应和练习效应既是可取的，也是不可避免的；他们可能会将在识别依赖关系、基数和瓶颈方面的学习经验迁移到后续任务中。我们采用了组间轮换的两组被试内设计，以避免携带效应。我们使用了涉及系统另一部分、但目标相同的不同问题

来限制对答案的记忆。为避免用户偏见，我们以中性的名称指代两项研究，使参与者不会预先猜测或将其与另一种应用方法联系起来。

此外，这两种可视化工具因其固有特性，在显示和交互方式上各不相同。二维工具缺乏交互性，而增强现实工具使用弹出窗口，这可能会产生影响。尽管如此，从参与者的反馈来看，在解决任务的难易程度方面并未观察到显著差异。这与研究的目标一致：评估增强现实方法能否在典型的微服务评估任务中胜过传统二维方法，尽管两者性质不同，且移动设备的尺寸较小，存在局限性。

研究过程面临兼容性威胁，导致部分参与者无法安装增强现实工具，因此，执行二维实验的参与者数量多于执行增强现实实验的参与者，人数分布不均。此外，参与者按照与收到研究材料相反的顺序执行研究，这影响了交叉实验在研究中的分布。然而，统计结果表明，先使用一种可视化方式完成任务，再使用另一种可视化方式完成一组相似任务，不会影响研究结果。参与者的移动设备也给增强现实相关研究带来了特定威胁，因为设备的屏幕尺寸不同，可能影响他们对系统的理解。然而，屏幕尺寸数据的分布不足，且缺乏足够的分类数据，因此无法将其纳入模型并考察其影响。

与参与者对流程和任务的理解有关的威胁已在反馈表单中予以考虑。大多数参与者表示内容清晰；然而，少数参与者认为，从他们的角度来看，任务不够清晰。这也可能是由于他们经验不足。因此，我们将经验水平作为本研究的主要因素之一。

最后，研究受参与者人数限制；我们选择优先为从业者提供有意义的结果，而不是扩大参与学生的群体规模。本研究采用被试内设计，使每位参与者都能使用两种工具，从而有效地将参与者人数从 20 人增加到 40 人。此外，参与者人数不足降低了部分测量结果的显著性，包括瓶颈检测测量结果。尽管如此，仍需通过更多参与者开展进一步研究，以确认其显著性。

3）外部效度：我们直接联系了联系人，从中选取参与者。不过，我们接触到了来自北美洲和欧洲、背景各异的广大从业者。我们调查了研究参与者以了解他们以往的经验，并相信他们的回答是可信的。最终，受试者样本相当多元，涵盖了微服务经验从初级到高级的不同水平。此外，我们通过简单随机方法分配研究任务，以避免结果出现偏差。

为避免参与者对所研究的工具和系统感到熟悉，我们考虑了首次向他们展示这些内容。我们使用了社区认可的第三方基准测试，以避免偏差。我们手动调整系统，以生成小型和大型系统规模的视图。不过，我们在参与者中随机分配了这两种变体。

4）结论有效性：为避免偏差，我们纳入了来自不同领域、具备不同专业知识的作者。研究结果通过自然文本和李克特量表答案收集意见，随后对这些数据进行了统计分析和描述性分析。此外，本研究的首要目标是回答三个研究问题，而用户研究还带来了关于实践意义、局限性和开放挑战的其他见解；在将增强现实工具应用于实践时，这些问题仍有待解决。

VIII. 结论与未来工作

我们开展了一项受控实验，比较参与者通过二维和增强现实（AR）呈现的服务调用图，对基于微服务的系统的主观可理解性。研究包含多项贴近实际的任务，用于衡量开发人员使用所研究工具理解系统的能力。研究重点考察了参与者的专业经验以及不同系统规模的影响。我们对任务答案和反馈进行了统计分析与描述性分析。

研究结果考察了参与者针对不同可视化方面的反馈，以及为提升系统可理解性和进一步开发增强现实系统应提供的信息。我们得出结论：增强现实可视化使新手开发人员能够像有经验的开发人员一样理解系统依赖关系。然而，有经验的开发人员使用增强现实分析部分任务时并未获得显著收益。其原因之一可能是，有经验的开发人员习惯于使用二维可视化评估系统，而引入增强现实可能影响了他们的表现。另一方面，增强现实在大型系统的缩放和可视化方面展现出显著优势；但对于小型系统的可视化，这种优势可能并不显著。

在未来的工作中，我们将仅针对资深开发人员重复这项研究，并分析工业系统，以更好地了解增强现实的实用性，以及开发人员理解系统所需的信息。我们也将考虑将交互式三维可视化作为一种替代方案。动态系统视图也能受益于先进的视觉视角。

致谢

本材料基于美国国家科学基金会第 1854049 号资助以及以下机构提供的资助所支持的工作：红帽研究院、乌拉·图奥米宁基金会（芬兰）提供的资助、芬兰科学院提供的资助（资助编号 349488 - MuFAno），以及项目 PGC2018-097265-B-I00。该项目由欧洲区域发展基金/西班牙科学与创新部—国家研究局资助。

参考文献

[1] F. Rademacher、S. Sachweh 和 A. Zündorf，“一种用于系统性重构基于微服务的软件系统架构的建模方法”，载于《企业、业务流程与信息系统建模》。Springer International Publishing，2020 年，第 311–326 页。[2] T. Cerny、A. Abdelfattah、V. Bushong、A. A. Maruf 和 D. Taibi，“微服务架构重构与可视化技术：综述”，载于《2022年IEEE面向服务的系统工程研讨会（SOSE）》，2022年。[3] M. E. Gortney、P. E. Harris、T. Cerny、A. A. Maruf、M. Bures、D. Taibi 和 P. Tisnovsky，“从动态视角可视化微服务架构：一项系统性映射研究”，IEEE Access，第 10 卷，第 119 999–120 012 页，2022年。[4] T. Cerny、A. Abdelfattah、V. Bushong、A. A. Maruf 和 D. Taibi，“Microvision：一种基于静态分析、用于在增强现实中可视化微服务的方法”，载于《2022年IEEE面向服务的系统工程研讨会（SOSE）》，2022年。[5] L. O’Brien、C. Stoermer 和 C. Verhoef，“软件架构重构：实践需求与当前方法”，2002年。[6] J. Bogner、J. Fritzsch、S. Wagner 和 A. Zimmermann，“微服务可演化性保障的行业实践与挑战”，Empirical Software Engineering，第 26， 5， 104, 卷，2021 [ ] 10 1007 10664 021 09999 26， 5， 104, 期，第2021 [ ] 10 1007 10664 021 09999 26， 5， 104, 页。在线。2021 [ ] 10 1007 10664 021 09999 可获取：https://doi.org/ 。/s - - - 9 [7] N. Alshuqayran、N. Ali 和 R. Evans，“迈向微服务架构恢复：一项实证研究”，载于《2018年IEEE软件架构国际会议（ICSA）》，2018年，第 47–4709 页。[8] G. Granchelli、M. Cardarelli、P. D. Francesco、I. Malavolta、L. Iovino 和 A. D. Salle，“迈向基于微服务的系统的软件架构恢复”，载于《2017年IEEE软件架构国际会议研讨会》，2017年，第 46–53 页。[9] J. Thalheim、A. Rodrigues、I. E. Akkus、P. Bhatotia、R. Chen、B. Viswanath、L. Jiao 和 C. Fetzer，“Sieve：从分布式系统的监控指标中获取可执行洞察”，载于第 18 届 ACM/IFIP/USENIX中间件会议论文集，丛书：Middleware ’17。美国纽约州纽约市：ACM，2017年，第 14–27 页。 [在线]。可获取：http://doi.acm.org/ 10 1145 3135974 3135977 。/ 。S. Esparrachiari、T. [10] Reilly 和 A. Rentz，“跟踪和控制微服务依赖关系”，Queue，第 16 卷，第 4 期，第 10:44–10:65 页，2018 年 8 月。 [在线]。可获取：http://doi.acm.org/10.1145/3277539。3277541[11] S. Ma、C. Fan、Y. Chuang、W. Lee、S. Lee 和 N. Hsueh，“使用服务依赖图分析和测试微服务”，载于《2018年第42届IEEE计算机软件与应用年会（COMPSAC）》，第 02 卷，2018年，第 81–86 页。[19] A. Schreiber、L. Nafeie、A. Baranowski、P. Seipel 和 M. Misiak，“在虚拟现实和增强现实中可视化软件架构”，载于《2019年IEEE航空航天会议》，2019年，第 1–12 页。[12] G. Parker、S. Kim、A. A. Maruf、T. Cerny、K. Frajtak、P. Tisnovsky 和 D. Taibi，“运行时微服务反模式的可视化：一项系统性映射研究”，IEEE Access，第11卷，第4434–4442页，2023年。[13] O. Group 等，“开放组架构框架 TOGAF——第8版”，企业版，2002年。[14] M. M. Lankhorst、H. A. Proper 和 H. Jonkers，“ArchiMate语言的剖析”，《国际信息系统建模与设计杂志》（IJISMD），第1卷，第1期，第1–32页，2010年。[15] B. Mayer 和 R. Weinreich，“用于微服务监控和管理的仪表板”，载于《2017年IEEE软件架构国际会议研讨会》（ICSAW），2017年，第66–69页。[16] H. Halpin、D. J. Zielinski、R. Brady 和 G. Kelly，“利用虚拟现实探索语义社交网络”，载于《语义网——ISWC 2008》，A. Sheth、S. Staab、M. Dean、M. Paolucci、D. Maynard、T. Finin 和 K. Thirunarayan 编。柏林，海德堡：Springer Berlin Heidelberg，2008年，第599–614页。[17] Z. Ma 和 Y. Bai，“基于虚拟现实的分布式系统监控工具”，载于《第2届计算机科学与应用工程国际会议论文集》，系列：CSAE ’18。美国纽约：计算机协会，2018年。[18] R. Oberhauser 和 C. Pogolski，“VR-EA：使用ArchiMate和业务流程模型和符号对企业架构模型进行虚拟现实可视化”，载于《业务建模与软件设计》，B. Shishkov 编。Cham：Springer International Publishing，2019年，第356卷，第170–187页，系列名称：《商业信息处理讲义》。[20] X. Zhou、X. Peng、T. Xie、J. Sun、C. Xu、C. Ji 和 W. Zhao，“面向软件工程研究的微服务系统基准测试”，载于《第40届国际软件工程会议论文集：伴随论文集》，ICSE 2018，瑞典哥德堡，2018年5月27日至6月3日，M. Chaudron、I. Crnkovic、M. Chechik 和 M. Harman 编。ACM，2018年，第323–324页。 [在线]。可访问：https://doi.org/10.1145/3183440.3194991[21] V. Bushong、D. Das 和 T. Cern`y，“使用静态分析重建微服务系统的整体架构”。载于CLOSER，2022年，第149–157页。[22] S. Vegas、C. Apa 和 N. Juristo，“软件工程实验中的交叉设计：优势与风险”，《IEEE软件工程汇刊》，第42卷，第2期，第120–135页，2016年。[23] J. Cohen，“统计功效入门。”2016年。[24] N. Cliff，“优势统计量：通过序数分析回答序数问题。”《心理学通报》，第114卷，第3期，第494页，1993年。[25] C. Wohlin、P. Runeson、M. Höst、M. Ohlsson、B. Regnell 和 A. Wess-lén，《软件工程实验：导论》，系列：Kluwer软件工程国际丛书。德国：Springer，2000年。