**集合数据的可视化和视觉分析：一项研究**

**摘要** 在过去的十年间，由于集合数据广泛的可用性和各个学科不断增长的可视化需求，集合可视化重要的发展被受到关注。从数据分析的角度可以看出很多集合可视化都集中在集合数据相同的方面，使用相似的数据聚合或者不确定建模方式。然而，对那些基本特性的反思和对工作的系统概括的缺失，使得可视化研究无法有效地鉴别新的或无法解决的问题和计划进一步的发展。在本篇论文中，我们从整体的角度分析并提供一个对集合可视化的概述。具体来说，我们研究了最近十年来的集合可视化工作，并从两个角度将它们分类。（1）他们建议的可视化技术（2）他们包含的可视化任务。从第一个角度，我们重点详细描述传统的可视化技术（例如表面，体积可视化技术）是怎么适用在集合数据，从第二个角度，我们重点详细描述可视化任务（例如比较，聚合）是怎么多样地展示集合数据。从集合可视化文献的研究中，我们还发现了一些研究趋势和一些未来的研究机会。

**关键词** 集合数据 可视化和可视化分析 文献研究 分类

**第一节**

**1.介绍和动机**

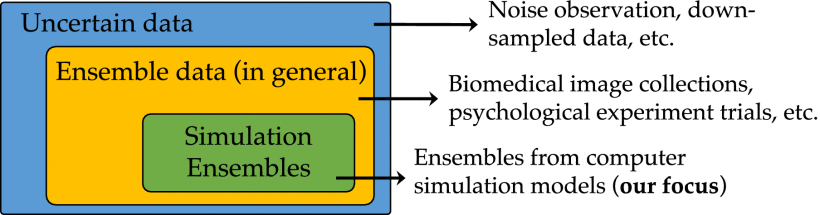
各个学科的科学家经常需要用复杂的电脑模拟模型来研究复杂的现实现象，这些模型通常被一套模型配置影响（例如输入参数值，边界/初始化条件，现象性模型等）。有特别配置模型的一次运行通常不足以了解物理现象。因为固有的不确定性，科学家经常需要尝试运行不同配置的模型来产生相同实验的多种结果。输出是时空结果的集合，通常称为集合。随着数据采集技术的发展和准确模拟各种物理现象增长的必然性（例如灾难管理，天气预测等等），越来越多的集合数据以前所未有的速度增长着。然而，由于集合数据复杂的本质（例如多维度，多价值），对于领域科学家来说是一个很重要的问题去完全理解集合数据。

在领导领域科学家去更好，更直觉地理解他们的数据的领域，可视化扮演着重要的角色。在过去的十年间，大量的集合可视化技术被介绍，令人鼓舞的结果被报道。例如，许多和可视化技术相关的轮廓和曲线被提议用来揭示集合数据空间特点的不确定性[3],[4],[5],[6],[7]。概率分布被广泛使用在集合可视化用来模拟潜在的不确定性[2], [8],[9],[10],[11],[12],[13]。不同类型的时间序列图和比较可视化方法已经被使用，用来捕获集合数据的趋势[14],[15],[16],[17],[18],[19]。为了有效地证明集合数据的不同方面，大量协调视图经常被使用，并和简便的交互连接一起用来便利用户的探索[20],[21], [22],[23],[24],[25]。

随着集合可视化出版物数量的不断增大，拥有一个当前最先进要求的构架概述变得越来越重要。这样一个结构视图将帮助我们提取这些作品的必要特征，进一步帮助我们鉴别当前的研究趋势和可能性的未来方向。将这个作为主要的目标，在这次工作中，我们展示一个在集合可视化领域的研究的结构化调查。

**1.1研究范围和目标受众**

从一个更广阔的角度来看，集合数据是不确定数据的一个子类[26],[27],[28],而不确定数据的不确定性的引入是由于相同物理量的大量实例。该类数据可以使用许多不同的的方式生成（和许多不同的学科）。在这次工作中，我们特别关注电脑模拟模型生成的集合数据（如图1所示）。除了集合模拟，也有很多不同来源的集合数据，例如生物图像收集[29];心理实验的重复实验[30];Snort警告和网络安全数据流量的收集[31]等。在这次调查中，我们只关注电脑模拟集合的原因如下：（1）这些集合已经覆盖了足够大的文献，并且众多学科领域科学家的主要关注点（2）电脑模拟通常和空间场有关联，并且场的每个位置的相同的物理属性的多个值的将模型的不确定性包括在内，因此对现存的可视化技术发起挑战。例如传统的体积绘制算法（对于确定性空间场）在面对集成模拟生成的多值空间场的可视化不能快捷适用。



**图一.**

该调查关注模拟模型的集合数据

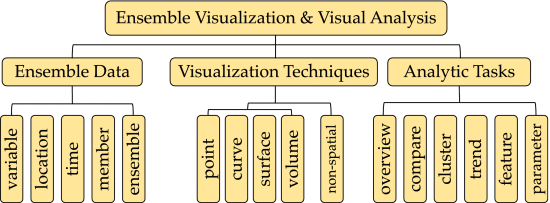
随着电脑模拟集合成为主要焦点，我们分析在过去10年在主要的可视化会议（IEEE可视化会议，欧洲可视化会议，IEEE空间可视化研讨会，IEEE大型数据分析可视化研讨会）和可视化期刊（IEEE可视化与计算机图形学汇刊、计算机图形学论坛、计算机图形学与应用）上的论文（2007年到2017年）。出现在这些论文的相关早期作品也被确定被包含进我们的调查。

本次调查的目标受众是那些已经拥有科学可视化背景和那些想了解更多不确定可视化，但还没有时间深入了解集合可视化的人。因此，我们将不会解释传统可视化算法的细节（像体积绘制如何工作），而是更多地集中在证明传统可视化算法是如何适用于集合数据。我们的研究也可以通过该领域的一个全面的概述，帮助现有的集合可视化研究者去更好地理解他们同事的作品。

和现存的相似的作品不同。从可视化文献中，我们发现有四部和我们类型的作品[32],[33],[34],[35]。我们作品和他们的区别在以下几个方面。第一，我们关注集合数据中更特别的组，即是电脑模拟集合。和Kehrer和Hauser所做的多方面数据的一般研究对比[34]，我们有限的调查范围帮助我们对集合可视化进行精确的分类，并对每个类别的作品用更好的细节进行详细的阐述。第二，对于集中的集合数据，我们的调查会更全面并包括更多方面。和Obermaier和Joy的短文相比[35],我们的调查不仅包含空间方面，而且包含其他方面，像集合数据的时间趋势和参数分析。和Koles ar等人的概念框架相比[32],我们的调查不仅讨论对比任务，还包括其他任务，例如集合数据的概述和聚类。第三，我们更关注在详细描述可视化技术和分析任务对集合数据是怎么样适用或者不同地展示的。这会限制我们作品的受众（对那些对可视化有一定知识的人），还会使我们的作品更有整体的特点。最后，和Love 等人2005年的作品相比[33]，他的作品总结了集合可视化三组方法（参数化，形状描述，运算符方法），我们的调查对这个主题提出了更现代的看法。

**1.2调查结构**

图2展示了本次调查的结构。这三个二级节点分别代表这本论文2，3，4章节涉及的主题。



**图2.**

本次调查的结构

我们通过回答以下问题来入手集合数据的基本概念：什么是集合数据；集合数据和传统科学数据的区别；在第二章中是什么造成集合数据可视化如此困难。简单易懂的数据描述是已经成形的，它识别集合数据的五个正交维度（即变量，位置，时间，成员和集合）。该描述是应用独立的，因此可以用来连接不同学科的集合数据。

第三章首先基于该技术是否适用空间数据对传统可视化技术进行划分。该划分和科学可视化技术与信息可视化技术的区别很像。即科学可视化技术更关注空间结构的数据，而信息可视化技术更关注数据的非空间描述。对于科学可视化技术，我们进一步划分为面向点，面向线，面向表面，面向体积的技术，并讨论在这四个组中传统技术是怎么样分别适用于集合数据。对于信息可视化技术，本次研究更关注它是怎么和科学可视化技术结合，通过协调多个视图展示集合数据的不同方面的方式，帮助视觉分析。这些技术经常出现在完整的可视化分析系统，这些系统是被科学家在某些特定条件下研发。我们在一个单独的章节总结这些技术。

集合可视化用于不确定分析的普通目标，这些分析实际上通常被固定的底层任务完成。第四章节会概况出六个常见任务，这些任务是我们从现存的文献中通过结构化分析得出的，即概述，对比，聚集，时间趋势分析，特征提取和参数分析。一些任务是集合数据所特有的，像是仿真因素的分析，然而其他任务是处理传统确定性数据的共享任务，像是比较和聚类。对于之后的任务，我们更关注详细阐述这些任务是不同地被展示用于适用集合数据的分析。

我们的调查展示集合可视化文献的结构化回顾，通过这些文献，我们已经确定了几个研究趋势和进一步的可视化挑战，我们已经在第五章节进行总结。

**第二节**

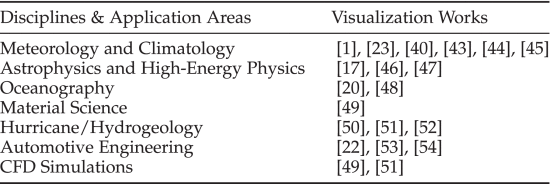
**2.插入数据**

集合数据（在我们调查范围内的）是指包含由电脑模拟模型生成的输出的集合的数据。这些输出既可以是相同模拟模型不同执行产生，这些模型以轻微不同的初始化条件或者扰动参数设置运行（Monte-Carlo仿真[1]）,或者由不同仿真模自身型的运行产生（多模型仿真[23],[36],[37]）。每一个在集合中的独立输出被称为集合成员或者实现[38]。值得一提的是相同集合的成员可能有相同的内部关系或者遵守着某些分布，这些分布是那些成员（例如使用统计模型来微扰模型参数）生成过程的结果。通常，集合数据通常生成是为了一个或多个以下的目的：

* 模型初始化/边界条件的判断或者校准[23],[39]
* 参数敏感调查[25],[39],[40],[41]
* 不确定性的量化/缓解[40],[42],[43]
* 模型对比或者模型缺陷研究[23],[37]

随着数据获取的重要进步和电脑能力的显著提升，这类型的数据被不可预测的速度[1],[2]和增长的多种的学科生成。表1列出了一些学科和相关可视化作品。

**表1** 生成集合数据的学科和应用领域



区分集合数据和传统确定性科学数据的的主要属性是他们额外的集合成员维度。一方面，集合数据有确定性科学数据集有的所有特性，例如多元，多维度，空间等等。另一方面，集合数据的各个成员加入新的维度进数据中。换句话说，一个集合数据可以被当成传统科学数

**2.1数据表示**

为了确定集合数据中的结构和逻辑模式，我们在本节中格式化他们的数据。本次展示也会帮助我们根据出版物中的集合数据的特征来组织出版物。

对于电脑仿真模型典型的执行，领域科学家需要设置N组仿真参数的数值，表示为：

P={p1,p2,...,pN}, (1)

每一个参数pi∈P都有一个兴趣范围[risi,riei]⊂R，它是基于参数的物理解释预先定义。来自这个N维参数空间的样本是参数向量xi¯∈RN，即

xi¯=(x1,x2,...,xN) （2）

表示为S的仿真模型将这个参数作为输入并生成一个时间序列S(xi¯)(一个集合成员mi)，该时间序列由T个时间步骤完成，即

S(xi¯)=mi={f1,f2,...,fT},（3）

其中fi是相应时间段的模拟字段。对于一些模拟，不同时间段的fi是相同的，然而其他模型可能有随时间变化的字段。

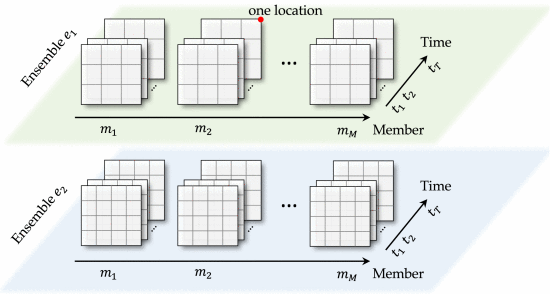
每一个字段fi由多个观察点组成（位置，参考方程式(4)）。根据模拟设置，这些位置可以被组织成规则网格（例如直角坐标网格，曲线网格[40],圆柱形网格[55]等等）或者不规则的网云[56]。

fi={l1,l2,...,lL} （4）

每个Li存储着多个标量/矢量变量的模拟结果（例如温度，速度）。变量组V，位置Li，可以被表示为：

Vli={v1,v2,...,vV}.

集合仿真由多个这样的执行组成，每一个使用一个不同的参数样本xi¯，或者一个不同的仿真模型。xi¯或者仿真模型的不同会导致集合成员的不同。在图片3的语义例中，集合数据包含M个成员；每个成员有T个时间段；每个时间段的空间仿真字段有16个位置(组成4x4的笛卡尔坐标系网格)；每个位置（展示为红点）记录多个变量。



**图3.**

集合数据的五个维度。此语义数据具有两个不同空间分辨率的集合，为e1和e2；每个集合有M个成员，每个成员有T个时间段；每个时间段的模拟字段在第一/第二集合中有16/25个位置；每个位置（红点）记录多个变量的值。

此外，我们发现领域科学家可能需要同时学习多个集合。这些集合通常紧密关联，但在关键的仿真配置中有不同。例如，为了捕获有趣的物理现象的的不同范围，气象学家会用不同的空间方案展示相同物理区域的气象模拟（使用气象研究和预测模型 即WRF）。通过对比结果的集合，他们可以研究不同的空间方案是怎么样影响输入参数的敏感性，执行的时间复杂度，仿真输出的准确性等。他们也可以获得更好的主意关于平衡计算成本和仿真准确性[40]。图三的语义例子展示了两个有不同空间方案的集合（一个是4x4的笛卡尔坐标系网格，一个是5x5的笛卡尔坐标系网格）。值得一提的是不同的集合通常使用不同的仿真参数组，并且集合的成员不一定相同[25],[40]。

从以上分析，我们确定了集合数据的五个正交维度，即变量，位置，时间，成员，和集合。我们将这个分析扩展到我们研究的所有论文的集合数据，并且发现这五个维度能很好地覆盖集合数据的空间。因此，我们根据那些论文的集合数据的五个维度对论文进行组织（表3的”集合数据”列）。一维的存在以为着该维度的大小大过一。例如一个时间变量集合有时间维度，因为它包含多个时间段；然而，时间维度不存在任何静态集合，因为时间段对于他们来说只有一个。

这五个维度也导致了集合数据的五个特征。即多元（由可变维度产生），多维[33]和时空[16],[19](由位置，时间维度产生)，多值[57](由成员维度产生)和分辨率[40](由集合维度产生)，这五个特征使得集合数据通常变得大和复杂。假设变量，观察点，时间段，成员和集合（在不同分辨率）分别是v,l,t,m和e，集合数据的空间复杂度将会是O(v⋅l⋅t⋅m⋅e)。所有五个数字可以变得非常大，所以集合数据的大小可以很容易到TB或者PB[57]。分析这类数据和提供变量关系，特征属性的不确定性，时间演变趋势，不同成员或者不同集合的变化的见解是很重要的。

参数数据和真值数据。集合数据经常会附带两个额外类型的数据。第一种是仿真模型的输入设定，它可以是边界条件，模型设置，参数设置等等。在这次作品中我们把这些可能的输入设定作为参数数据。对于集合可视化，在输入参数和输出集合之间建立联系是最重要的分析任务之一。第二个类型是真值数据。例如，从卫星观察得到的天气条件可以当成WRF集合的真值[25],[40]。真值数据经常被使用来评估集合成员的质量和校正模型参数[39]。参数数据(仿真输入)通常和集合数据（仿真输出）相关联，尽管在某些情况下他们可能表示分析的重点。然而，真值数据的存在高度依赖领域问题。

**2.2出现的可视化问题**

在集合数据的正式定义和全面解释之后，因为传统可视化技术不能直接应用于集合数据，我们想指出两个出现的问题。

额外维度。集合数据引入了一个新的数据维度，即成员维度，而传统可视化技术通常不包含它。例如，直接体积渲染算法可以渲染单值标量场体积；但它不能渲染多值体积（每个网格点有多个标量值），这是因为在多值字段中采样和插值不能被很好地定义[58]。该算法对集合数据的直接适应是在可视化之前，将一个网格点的一组值合并成一个数值，例如平均体积的体积渲染。然而，该适应通常是不适当的。例如，如果底层的集合成员（每个网格点的数值集合）遵守着双峰或多峰分布，平均体积的可视化将会被误导。

多面性。集合数据的五个维度使它成为多面性的[34]，并且领域科学家经常需要同时研究这些方面。然而，一个单独的可视化技术只能覆盖五个维度的一个或两个。例如，如果用一个平行轴用于一个变量，一个折线图用于一个集合成员，那么一个平行坐标轴可以覆盖集合数据的成员和变量维度。然而，每个成员的的时空不能被同时显示。集成可视化分析系统（在领域科学家的特定需求下开发）经常是有人需求的。这些系统(像Ovis[20], Noodles[43])具有多个协调视图和友好交互设计，已经成功帮助领域科学家解决集合数据中的实际现实问题。

我们的研究回顾了传统可视化技术是怎么样处理集合数据的额外维度（第三节），和不同的分析任务是怎么样演示和连接集合数据的多面性(第四节)。

**第三节**

**可视化技术**

电脑仿真集合通常嵌套在具体的空间场/领域，并且空间场的可视化是领域科学家的基本要求之一。因此，我们根据空间数据几何特征将可视化技术分为四个类别，即点指定，曲线指定，表面指定和体积指定。

由于额外的成员维度，在这四个类别的传统可视化技术通常不能直接适用于集合数据。集合可视化技术通常采用统计聚合（在可视化前），可视化合并（在可视化后），或者两者结合来处理额外的成员维度。之后，可以引出一个非常通用的集合可视化流程，如图4所示。



**图4.**

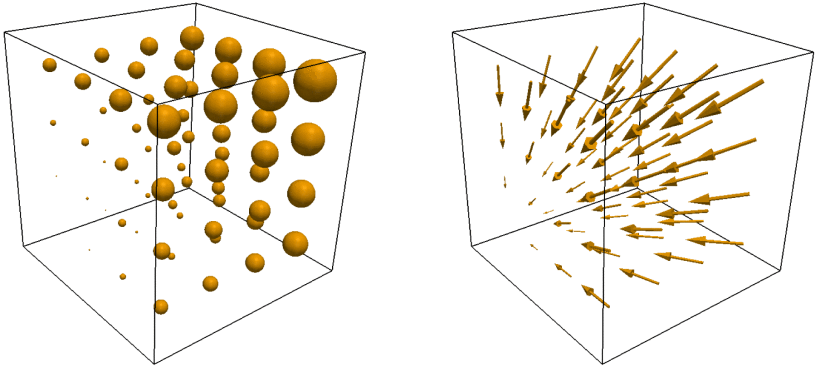
普通的可视化途径。集合数据将经历：可视化之前的统计聚合步骤；或者可视化之后的视觉合成步骤，或者两个步骤的结合。聚集和合成步骤在带有虚线边框的矩形中，表示它们可能并不总是出现在该途径的所有路径中。

聚合步骤使用统计摘要来聚合成员维度。可视化可以被适用在结果摘要上，用于揭露集合数据的集体模式/行为。统计摘要的例子包括：均值计算，变化幅度，集合数据的顺序统计[6],[7],对集合成员的概率分布的建模[8],[12],聚集集合成员[4],[5]等等。注意，统计聚类可以适用于不同成员的物理量（例如计算不同成员中记录的温度值的均值）或者提取出他们的几何特征。例如，集合成员可以表示为等值面或者曲线（既可以从集合仿真中输出直接获得，也可以从预处理步骤中产生）。它们的聚类（例如平均等值面，聚类曲线）通常需要更多复杂的计算。可视化之后处理成员维度的步骤会将单独成员的可视化（或者不同统计摘要的可视化）进行合成，从而揭露出集合数据的集体行为。典型的视觉组成方法包括：叠加，并置，嵌套和重载，它们是被Javed和Elmqvist调查得出[60]。我们在3.1,3.2, 3.3和 3.4节中，根据这条总体流程的不同途径（例如可视化之前的聚类，可视化之后的组合）,来讨论传统可视化技术的四个类别是怎么样适用于集合数据。

电脑仿真集合也可以被集合/提取为非空间数据以便概述。为了可视化有不同等级细节的集合数据，不同类型的非空间可视化技术和空间可视化技术一起被使用。我们在3.5节会讨论这些技术。

**3.1面向点技术**

面向点可视化技术映射每个网格点的数据(空间场中的)到选择的可视化通道。应用到确定性科学数据的面向点技术的例子包括假染色(将数值映射到颜色)和图形可视化(将数值映射到形状)[61]。图5展示了一个图形可视化的例子。在左侧，球的大小编码着相对于位置的标量值，在右侧，箭头的大小/方向编码着标量的大小/方向。



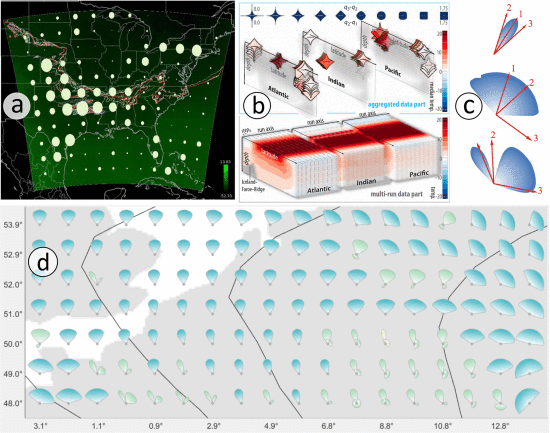
**图5.**

图形可视化。左侧：使用球的大小编码每个网格点的标量值；右侧：使用箭头的大小/方向编码每个网格点的矢量的大小/方向。

对于集合数据，每个网格点都有多个标量或者矢量值，并且所需的可视化应该能够有效地演示那些数值的集合模式。按照图4的总体流程，面向点的技术已经适用于集合数据:(1)合计每个网格点的多个值，并且用传统方法可视化合计结果（2）先可视化单个值，然后合成每个网格点的可视化。

**3.1.1可视化之前的聚合**

第一组面向点的可视化技术合并每个网格点的数据，并用图形去描述合并结果。例如，Sanyal 等人[43]计算每个网格点的95%置信区间的数值，并将该区间映射到圆形的大小(图6a)。Kehrer等人[61]计算每个网格点的集合成员的顺序统计量，并用超椭圆形展示统计量(图形图例在图6b的顶部展示)。颜色，总大小，超椭圆的上部形状和下部形状分别代表着中位数，四分位数范围(即q3-q1),q3-q2的范围，q2-q1的范围(q1,q2,q3代表着每个网格点的25%，50%,75%的集合数据)。



**图6.**

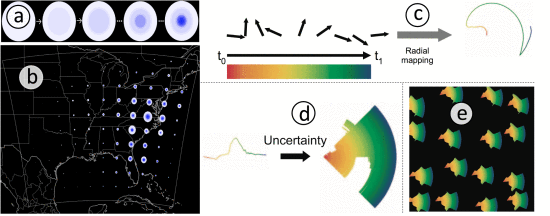
可视化之前的聚合：（a）使用圆形图形显示每个网格点的95％置信区间，图片由Sanyal等人提供。 [43]©2010 IEEE。 （b）超椭圆图形的颜色，大小，上下形状编码所有成员的中位数，四分位数范围和亚四分位数范围，图片由Kehrer等人提供。 [62]©2011 IEEE。 （c）使用小叶图形显示每个网格点上矢量方向的形式；（d）在仿真域中放置小叶图形的结果。 图片由Jarema等人提供。 [63]©2015 IEEE。

对于矢量场集合，Jarema 等人[63]用小叶图形去编码每个网格点矢量方向的形态。正如图6c所示，顶端的图形代表着变化不大的矢量方向的单值分布；中间的图形也代表着单值分布，但有大的变化；底部的图形代表着双峰分布。图6d展示了在整个模拟字段放置小叶图形后的可视化结果。

合计统计也可以编码到网格点的位置(例如将网格点转移到新的地方，该地方中有相似数据集的点会更加靠近[64])，网格点的时间行为(例如利用闪烁去编码变化[17])等。完整的例子在表3找到。

**3.1.2可视化之后的组成**

第二组的面向点的可视化技术首先可视化每个集合成员，然后合成同一个网格点中不同成员的可视化。例如，Sanyal等人[43]提出分级不确定性图形，在该图形中每个成员首先被可视化为圆形图形。圆形的大小表示相对于的集合数值和平均值的不同。同一个位置的那些的圆（代表不同的成员）被覆盖/组合是为了证明位置的不确定性，正如图7a所示（为了避免遮盖，将有深颜色的小圆放在浅颜色的大圆上面，即圆按照他们的大小排序）。图7b展示了使用分级不确定性图形去可视化整个字段的结果。Hlawatsch 等人[51]提出了针对不确定流场的流量雷达图形。对于他们的数据，每一个网格点都有随时变矢量集合。他们首先通过径向映射可视化每一个时变矢量为曲线状图形（图7c），然后合并所有成员的曲线状图形得到一个雷达图形（图7d）。图7e展示了在不同位置的模拟场中放置流量雷达图形的结果。



**图7.**

可视化后的组合。（a）分层次的不确定性图形：代表着单个成员的同心圆被有序重叠在一起；（b）在模拟场中的分层次不确定性图形，图片由Sanyal等人提供。[43]©2010 IEEE。（c）流量雷达图：通过径向映射可视化每个时变矢量；（d）通过视觉来编码成员的不确定性； （e）将流量雷达图应用于模拟场。 图片由Hlawatsch等提供。 [51]©2011 IEEE。

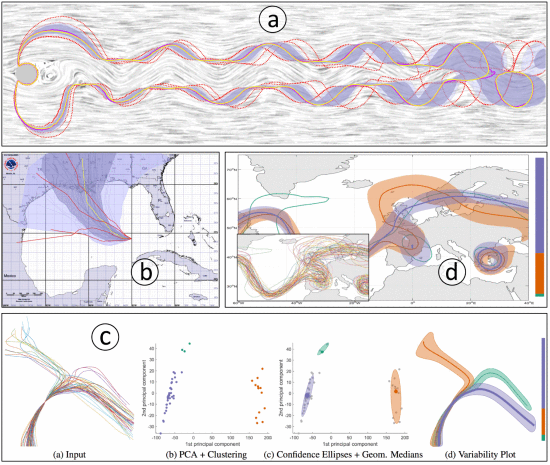
**3.2面向曲线技术**

面向曲线可视化技术从空间仿真领域提取并可视化曲线状特征。传统的面向曲线技术包括2D标量场的等值线可视化，2D/3D矢量场的流线和路劲可视化。给定一个确定性的标量场（或者确定性标量场的粒子位置）的等值线，可以用确定性的算法提取出曲线状特征（例如水平集提取方法，Runge-Kutta积分方法）。

对于集合数据，一个等值（一个粒子位置）对应着多个曲线状特征（每个特征来自一个集合成员），并且这些特征反应出编码进集合数据的不确定性。根据一般的集合可视化途径（图4），面向曲线技术已经适用在集合数据中，通过以下方法：（1）合并曲线和统计摘要，并可视化结果摘要；（2）首先可视化单个曲线（集合成员），然后合成曲线来揭露集合行为。

**3.2.1可视化之前的聚合**

第一组面向曲线集合可视化技术首先获得曲线状特征的概要统计（例如顺序统计，概率分布，聚类），然后将这些统计信息用可视化进行编码。例如，轮廓箱图[7]基于数据深度的概念得出曲线的顺序统计信息，并可视化中位数（图8a的黄色曲线），50%的间隔（深蓝色带），100%的间隔（浅蓝色带）和异常值(红色虚曲线)，以显示具有不确定性的等高线集合。曲线箱图将等高线箱线图的概念延伸到流线和路径图。图8b展示了用曲线箱图进行路径集合（50个仿真飓风轨迹）的可视化。除了异常值用红色实曲线展示，都是采用相同的配色方案。



**图8.**

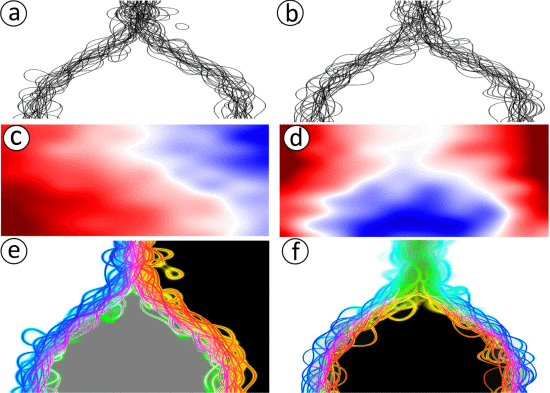
可视化前的聚合。（a）轮廓箱型图：计算集合曲线的顺序统计信息和可视化统计信息。图片由Whitaker等人提供。[7]©2013 IEEE。（b）曲线箱线图：将轮廓箱线图扩展到流线图/路径图。图片由Mirzargar等人提供。 [6]©2014 IEEE。（c）流线变化图：首先将集合曲线离散为高维矢量；然后将矢量投射为2D并且聚类成三组；映射每组的几何中位数和置信椭圆回曲线空间为变化图。 图片由Ferstl等提供。 [5]©2016 IEEE。（d）将变化图的概念扩展到轮廓。图片由Ferstl等提供。 [4]©2016 John Wiley＆Sons，Inc。

曲线状特征也可以表示成高维向量（通过将曲线离散成一系列不连续的点）。Ferstl等人提出的流线变化图[5],该图使用PAC算法将集合曲线投射为2D点，并且将它们聚集在2D空间的不同组。正如图8c所示，用三个波段/波瓣（蓝色，橙色和绿色）可视化三组集合曲线。每个波瓣中间的曲线和每个波瓣的形状分别反映着2D空间中每个曲线的的几何中值和置信椭圆。每组的中值曲线的厚度也编码着在那组中集合曲线的数量。将该思路延伸到轮廓，Ferstl等人也提出了轮廓变化图[4]，正如图8d所示。

值得一提的是，一些面向曲线的2d技术也可以适用（或扩展到）于3d矢量场，像轮廓箱图[65]和轮廓变化图[4]。因此，这些技术既属于面向曲线（用于2D矢量场）和面向表面（用于3D矢量场）。表3中可以找到详细的关于我们是怎么样分类这些作品。

**3.2.2可视化之后的组成**

第二组面向曲线集合可视化技术首先可视化单个曲线，然后把合成曲线（从不同集合成员提取）在一起来可视化整个集合。意大利面图（该图是用于2D轮廓集合）是该组最直观的例子，在图9a所展示。不同的颜色也可以被分配给不同的曲线来区别不同的集合成员[23]。



**图9.**

可视化后的组成。（a，b）用两个意大利面图可视化两个轮廓集合，可视化揭露了两个集合的差异很小；（c，d）分别来自集合（a，b）的一个成员；从小到大的数值映射为从蓝色，到白色（等于等值），再到红色；（e，f）编码轮廓分布和拓扑结构进意大利面图。图片取自Pfaffelmoser和Westermann的作品[66]©2013欧洲图形协会。

这组技术可以用非常直观的可视化来演示集合数据。然而，当集合成员的数量很多时，可视化很容易变得杂乱无章，甚至产生误导。例如，图9a和图9b有两个意大利面图的两个不同集合。总体而言，这两个集合似乎有非常相似的轮廓形状和轮廓分布，但这不是真实情况。图9c和图9d分别从图9a和图9b的集合演示中可视化两个单独的成员（在颜色映射矢量场中）。从小到大的数据值被映射成颜色，分别为蓝色，白色（等于等值）和红色。显而易见的是，图9c和图9d的轮廓形状非常不同。这两个意大利面图合成在每一个集合的轮廓一起，所以不能揭露不同性。Pfaffelmoser和Westermann通过三个颜色方案揭露了轮廓方向（从子集到父集的方向）和梯度分布的方式解决了这个问题，这三个颜色一个是是用在背景两个用在前景。背景颜色方案（炭黑色：[白色-灰色-黑色]）表示了轮廓的方向。对于黑色区域（图9e，9f），所有成员的值都小于等值；然而，对于白色区域，所有成员的值都大于等值。灰色区域强调了空间位置，该区域成员的值即有大于等值也有小于等值的。该配色方案有效强调了相似意大利面图中轮廓方向的不同。两个前景配色方案，CI：[黄色-绿色-青色]和Cu[红色-品红色-蓝色]，分别表示着各个等值线低和高的梯度。Cu的存在表示沿着等值线的梯度大小恒定不变，而高低梯度区域之间没有对比度。 同时Cl的存在表示梯度强度低得多。最后的意大利面图可视化（图9e，9f）可以有效地区分有不同曲线方向和梯度分布的两个集合。

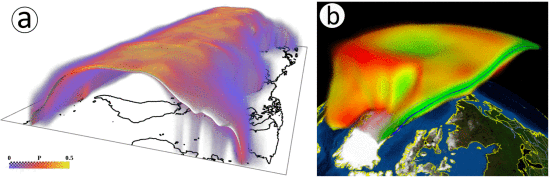
**3.3面向表面技术**

面向表面可视化技术从3D标量/矢量场中提取表面结构（例如等值面，流表面）进行可视化。传统面向表面技术的例子包括移动立方体法，等值面体积渲染（基于射线投射算法）和流表面可视化。

对于集合数据，多个表面结构同时存在，这使得传统表面提取和可视化技术不容易应用。可视化关键的挑战是同时解决多个表面并有效地这些表面的潜在的不确定性。从我们研究的可视化文献来看，面向表面技术已经通过以下方式适用于集合数据：（1）计算不确定模拟场的表面统计信息（例如每个体素中的曲面交叉概率），并用体积可视化技术可视化统计信息（2）合成各个成员的表面可视化结果并解决结果的可视化混乱和遮挡问题。

**3.3.1可视化之前的聚合**

第一组面向表面集合可视化技术首先执行在表面上的统计聚合，然后可视化所有表面的聚合的统计信息。例如，Pöthkow和Hege[68]提出了等高线位置的不确定性，该方法将集合信息聚集在概率场中，并使用体积渲染去可视化该概率场。对于感兴趣的等值面，（概率场中）体素中的每个概率值是水平交 叉概率（LCP）值（即等值面通过该体素的概率）。具体来说，他们首先将每个网格点中所有集合成员的数据建模成概率密度函数（PDF）。通过采用基于蒙特卡洛的采样策略为每个体素生成多个实现/采样来计算LCP值。在生成的样本中，一个单元格的数量意味着穿过单元的等值面构成该体素的LCP值。生成的LCP体积展示了等值面通过该概率场的概率，可以通过体积可视化方式进行可视化。之后，他们通过增加PDF体积和提出的概率移动立方体的相关性来扩展他们的工作[12],正如图10a所示。左下角的图例显示了映射LCP值（从0到0.5）到可视化通道（透明度和颜色）的传递函数。



**图10.**

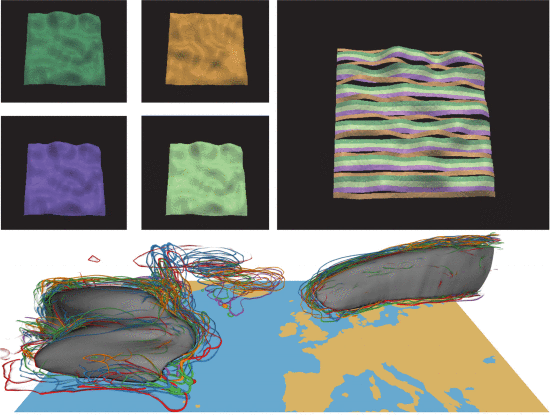
可视化之前的聚合。 （a）概率行进立方体使用直接体积渲染可视化特定等值的平交概率体积。 图片由Pöthkow人等提供。 [12]©2011 John Wiley＆Sons，Inc.（b）等值面交叉概率的体积渲染视图，其中概率计算与射线投射算法集成在一起。 等值面对应于某个大气温度值，并且距平均表面的距离用绿色（低）到红色（高）进行颜色编码。 图片由Pfaffelmoser等人提供。 [72]©2011 John Wiley＆Sons，Inc.。

针对该方向继续研究，更多关于旨在精确建模每个网格位置的不确定性（即PDF）的工作被提出。这些工作应用参数[12],[68],非参数[69]和两者混合（参数和非参数[13]）分布来模拟变化集合数据集中的位置不确定性。另外，除了蒙特卡洛采样法外，也提出了闭式解来计算分布场（PDF体积）中的概率等值面[70][71]。

该组的另一个例子是Pfaffelmoser等人的作品[72],正如图10b所示。他们提出了另一个计算首次穿越概率的方法（即第一个通过射线穿越等值面的概率），该方法可以集成进射线投射算法中，从而可视化有空间相关性的概率的等值线。他们介绍了一个增量更新的方法，该方法可以有效地将概率计算整合进从前到后的体积射线投射。为了有效地对位置和几何不确定性可视化，他们提出了一个新的颜色映射方案，该方案基于可能性的表面点与平均表面的近似空间偏差。图10a展示了他们关于不确定性等值面（大气温度）的可视化，根据颜色编码，其中到平均表面（蓝色曲线）的距离是从绿色（低）到红色（高）。

**3.3.2可视化之后的组成**

第二组面向表面可视化技术首先可视化各个集合成员（作为表面），然后视觉上合并这些成员并解决潜在的遮挡问题。例如，Alabi等人提出了集合表面切片，该方法将各个表面切分成多个切片，并将来自互相靠近的不同成员的切片组织在一起，从而进行成员结构对比。正如图11（上面）所示，四个集合表面（左边）在一个视图被摆放在一起（右边：每一个都以相同的颜色被展示成一系列不连续的表面切片）。从合成的可视化中的亮度不连续看，可以容易地识别出不同于其他三个的橙色成员/表面。该组中另一个例子是Demir等人的集合可视化作品[74](即屏幕空间轮廓)。他们使用表面轮廓（在某个视觉方向）来可视化3D表面（从不同集合成员中提取），从而缓解遮挡问题。正如图11所示（底部），在相同的3D屏幕中，不同表面的轮廓被聚类（用不同颜色编码）结合在一起。灰色表面展示了所有成员的平均表面，并且作为空间上下文进行工作和可视化所有表面的总体趋势。



**图11。**

可视化后的组成。上面：集合表面切片，四个集合表面被叠加在一起。用户可以快速确定从亮度不连续方面，橙色表面不同于其他三个。图片由Alabi等人提供。 [73]©2012 SPIE。下面：屏幕空间轮廓，灰色等值面代表着平均曲面，集合成员可视化为等值面的轮廓。 图片由Demir等人提供。 [74]©2016 IEEE。

**3.4面向体积技术**

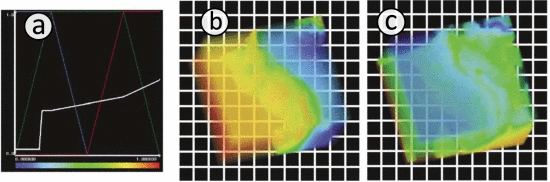
面向体积可视化技术旨在演示3D仿真领域，并揭露该领域重要的内部结构。直接

体积渲染是最普通的传统可视化技术，它通过射线投射算法将3D确定性矢量场可视化为2D图像。

一个集合数据集通常有多值体积（也可以被认为是单值体积的集合），这使得传统体积可视化技术不能直接应用。遵循一般流程，我们发现体积可视化技术通过合并多值体积为单值体积或者PDF体积，然后应用传统可视化技术到集合结果上的方式，已经适用于集合数据。进一步用来合成分离可视化结果的组成步骤也可能存在。

**3.4.1可视化之前的聚合**

第一组面向体积集合可视化技术通过合并每个网格点的多个例子为单个例子的方式，合并多值体积为单值体积。例如，假设每个网格点的集合数据都遵循着高斯分布，每个网格点的数值相对于的中值和标准偏差可以有效表示着多值体积。换句话说，所有体积的集合可以合并成两个体积（一个平均体积和一个标准偏差体积）。这两个体积可以成为直接体积渲染的输入。Djurcilov等人[59]通过一维传递函数，将合并平均和不确定数值（例如标准差和方差）编码成可视化物的颜色和透明度。正如图12a所示，白色曲线描绘不确定性映射（不确定性较高的区域的透明度较低）；底部的颜色带代表着颜色映射。图12b和图12c展示了变量“盐度”和“温度”（海洋集合中）的可视化结果。

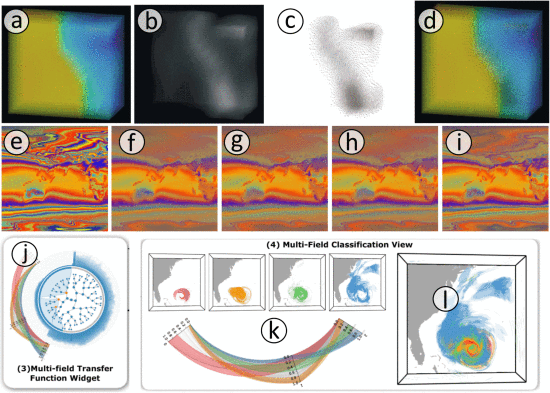


**图12。**

可视化前的聚合：（a）一维传递函数将平均值映射为颜色，和不确定性映射为透明度。（b，c）盐度和温度体积的体积绘制结果。 图片由Djurcilov等人提供。 [59]©2002 Elsevier Science Ltd.。

**3.4.2可视化之后的组成**

集合体积，即平均值和不确定体积，也可以分布渲染并通过合并步骤可视化在一起。例如，Djurcilov等人[59]通过单独渲染可视化海洋体积的平均值和不确定体积。图13a和图13b是平均平均值和不确定性体积的可视化结果(两个图像展示了相同视觉方向的两个体积)。不确定性体积被渲染成灰色大图像。他们反转了图像的像素值，并将图像抖动为灰度图。抖动尝试将黑色点（代表不确定性）平均分配在图像的适当位置，正如图13c所示。最终视觉合成步骤将平均体积（图13a）和不确定体积（图12c的灰度图）的渲染结果叠加在一起，从而将不确定性信息和平均体积一起传递，正如图13d所示。



**图13.**

聚合和组成：（a，b）海洋集合的均值和不确定体积的体积绘制结果；（c）反转不确定性渲染结果的像素颜色，并将其抖动为灰度图； （d）（a）和（c）的合成结果。 图片由Djurcilov等人提供。 [59]©2002 Elsevier ScienceLtd。使用基于体积可视化的GMM渲染空气温度集合的结果：（e）平均体积可视化； （f-i）使用一，二，四和六个高斯组成的GMM体积可视化。 图片由Liu等人提供。 [9]©2012 IEEE。 （j）飓风数据集的距离似然树； （k）绘制不同范围似然体积的结果； （l）将（k）中四个范围似然体积的渲染结果合成在一起。 图片由He等人提供。 [75]©2017 IEEE。

在某些情况下可视化平均体积是有误导性的，因为底层集合成员可能不遵循高斯分布。Liu等人[9]将每个网格点的集合成员建模成高斯混合模型(GMM)。结果，原本的集合体积变成了GMM的一个体积。然后他们可以对GMM体积进行采样，重新构造单值标量体积并用体积渲染进行可视化。图13f，13g，13h和13i分别展示了和平均体积对比（图13e），使用一个，两个，四个，六个高斯分量的渲染结果。在一个固定的视觉方向，反复从GMM中进行采样和可视化采样的单值体积可以产生一系列渲染结果（2D图像）。以动画的形式逐帧播放这些图像，会在不同图像区域产生闪烁效果，并且闪烁的频率代表着这些区域的不确定性等级。动画可以被当作一系列2D图像的组成，并且合成是随着时间维度进行的。

在该类中较近的一个例子是He等人的作品[75]。他们将每个网格点的集合成员建模成PDF，并且将PDF平均分成不同的数值范围。然后每个网格点都有一组用于标量的原始集合的“范围似然值”。例如，假设原始集合有100个成员，所有成员都有从0到255的标量值。每个网格点的PDF被分成四个平均的数值范围，范围1（0-63），范围2（64-127），范围3（128-191），范围4（192-255）。原始100个体积可以被合并进四个范围似然体积，即范围1、2、3和4的似然体积。范围似然体积的每个网格点都记录着在对应范围内标量值的概率。He等人将范围似然体积聚类并用分层聚类的方式建立一个这些体积的范围似然树（如图13j所示）。然后他们可以探索树并用不同范围似然值进行渲染（图13k）。不同范围似然值的渲染结果可以用多元体积渲染技术（用多元高斯传递函数）[76]进行合并。（图13l）

**3.5 非空间数据的可视化技术**

非空间数据的可视化技术也已经被广泛使用在集合可视化。绝大多数的技术可以不用任何改变轻易地应用于集合数据。例如，Hazarika等人[76]使用平行坐标图去展示集合表面的顺序统计量。Wang等人[25]使用树状图和热图来展示在多分辨率天气集合中成员相似度的概述。Bock等人[46]结合散点图和图形可视化来提供一个全局概述和空间天气集合成员的对比。Demir等人[76]采用折线图和条形图（即多图表）来演示集合成员的数值分布。

非空间可视化技术大多数是为了演示集合数据的不同等级的抽象/概述，并且他们经常和和大量其他空间/非空间可视化技术一起使用。这些技术通过不同可视化视图一起演示集合数据的不同方面。有这些结合在一起作为一个集合视觉分析系统，领域科学家可以执行特别的视觉分析任务和实现他们的分析目标。我们将在第4节讨论这些任务的细节。

**第四节**

**分析任务**

集合可视化的一般目标是描述集合数据的不确定性编码，从中推测有用的信息。实际上，一般目标通常具体的视觉分析任务完成。从集合可视化作品中，我们得出六个任务，这些任务覆盖了集合可视化文献的主要部分。请注意，这些任务不仅仅只是为集合可视化设计的，也可以被运用在其他类型的可视化工作。在本次调查中，我们关注这些任务是怎么样对集合数据进行处理的（即它们是怎么覆盖成员/集合维度）。这六个任务是：

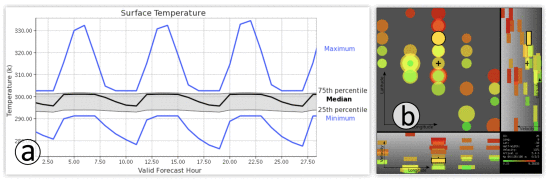
1. 概述：提供集合数据的简洁视觉总结和传达总体不确定性。
2. 比较:使用并置，叠加和显示编码来视觉识别两个成员或两个集合（即在两个空间位置的两个集合值）的区别。
3. 聚类：将成员或集合分割成单独的组，其中相似的成员或集合在相同的组。
4. 时间趋势分析：揭露一个成员，一些成员，或全部成员是怎么样随时间演变的。
5. 特征提取：从集合数据的不确定字段中提取几何特征（例如涡旋，涡流）或者拓扑特征（例如源，鞍形，汇点）。
6. 参数分析：在集合数据和模拟参数之间建立联系。

在可视化领域，这些任务似乎是最受欢迎的任务，并且它们反应领域科学家最核心的兴趣和需求。然而，值得一提的是它们不是完整的任务组，并且在这次调查中我们的任务不是找出所有可能的任务。并且，这些任务不是相互正交。一些任务在某种程度上依赖其他任务。例如，为了建立集合数据和模拟参数的联系，经常需要分别聚类集合数据和模拟数据，并从两方面对聚类进行关联。在这里，聚类和参数分析任务被执行，并且参数分析任务依赖于聚类结果。

**4.1 概述**

概述旨在展示数据的高水平视觉抽象。它通常被设计在任何交互之前可以被观看，因此，概述经常作为一个静态场景被展示，并为接下来的探索提供丰富的提示。

对于集合数据，概述任务是为了展示所有集合成员和揭露他们的集体行为（通常是空间或者时间行为）。为了使可视化表示简洁，抽象的某些级别可能需要应用到集合成员，例如，基于方法（即图8a，8b的轮廓线箱图[7]，曲线箱线图[6]）的箱线图通过顺序统计信息展示曲线/轮廓（即成员）集合的概述。从统计信息中，可以确定整个集合的主要的趋势，规模差异，离群值。时间序列图将每个集合成员抽象/集合在一起，并将其显示为一条线。展示代表不同成员的线并在2D时间线图（一个坐标轴是时间，另一个是感兴趣数量）表中叠加展示，从而演示一个集合整体的时间趋势[23][46]。这些线（代表单个成员）可以进一步合并来避免视觉混乱。例如，Potter等人[23]用分位趋势图（图14a）演示集合数据的时间概述。最小和最大边界（蓝色曲线）展示了整个集合的数值范围；分位数带（灰色曲线）显示了在集成中心排序50％的成员；中位线（黑色曲线）展示了所有成员的代表趋势。



**图14.**

（a）分位数趋势图：计算时间线图中表示不同集合成员的线的顺序统计，并可视化其顺序统计以便于集合概览。 图片由Potter等人提供。 [23]©2009 IEEE。 （b）参数空间概述的图形：颜色和大小分别表示速度和打开角度（两个参数）。 每个图形的位置反映了相应集合成员的经度，纬度（另外两个参数）。 图片由Bock等人提供。 [46]©2015 IEEE。

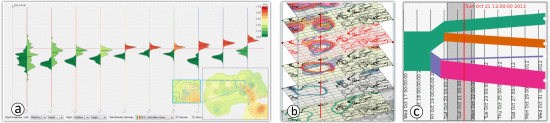
除了前面提到的空间概述或者时间概述，所有集合成员的参数设定概述对于领域科学家至关重要的。例如，Bock等人[46]将和单个成员关联的四个参数（经度，纬度，初始速度和打开角度）编码成圆形图形，并且在2D图中展示所有成员的图形，从而概述集合参数设定。正如图14b所示，2D图被分割成三个子图，子图展示参数空间的三个切口。左上角视图展示每个成员的经度和纬度坐标，而左下角和右上角视图展示经度和纬度方向的速度。每个图形的颜色和大小是对应集合成员的速度大小和打开角度对应的编码。从所有参数设置的概述中可以看出，用户可以选择单个图形从而进一步探索对应的集合成员。

在某种程度上，大多数集合可视化作品提供了所有集合成员的概述，因为它是非常基础的任务。在表3的“概述”列，我们使用字符S，T和P（表示空间，时间和参数概述）来反映在个人任务中概述任务关注的方面。

**4.2比较**

比较，作为通用的视觉分析方法，允许我们确定不同数据例子的相似点/不同点。Gleicher等人[79]通过他们对可视化文献的分析，将比较可视化技术分类成三个组：重叠，并置，和显示编码。例如，对应集合可视化，该分类依然适用，但对比的对象不仅限制为俩个数据例子（即两个成员/时间段），还包括例子的两个集合（即两个数据/曲线的集合）[80]。我们给出了三个比较技术中三个分类的两个情况的例子（集合可视化作品）。

并置为了对比将多个可视化结果并排放置。它依赖用户的短期记忆，通过频繁切换视觉关注点来确定并排视图的不同。例如，Demir等人[78]在3D体积集合中，通过并排放置单个体积/成员来对比，展示并列对比。选择体积中特别范围的值（使用多图表界面），将会突出这些体积中不同区域。通过对比，用户可以感知在不同3D体积（即集合成员）中数值是如何分布的。Höllt等人[20]对两个不同位置的两个值的集/集合在时间上进行并置比较，以揭示两个位置的不确定性和不确定性演变。正如图15a所示，每个垂直轴展示两个位置在一个时间段中的比较（标志在右下角插图中）。在两个位置中数值的两个集合作为两个PDF进行建模，并放在坐标轴的两侧进行并列比较（在我们时间趋势任务中，会在4.4节对插图进行进一步的解释）。



**图15.**

演示时间趋势。（a）用小提琴图形可视化和对比两点的时间趋势。图表的水平/垂直轴是时间/海平面。每个时间戳，两点的集合成员的PDF显示为两个并排（用作对比）的小提琴图。图片由Höllt等人提供。 [20]©2014 IEEE。 （b）具有堆积轮廓变化图的集合曲线的时间趋势； （c）平行集代表着集群分裂趋势。图片由Ferstl等人提供。 [15]©2017 IEEE。

叠加对齐对象，使它们可以在相同上下文进行对比，从而避免潜在的心理认知，以便有利于它们的对比。例如，Alabi等人[73]将四个表面进行叠加（正如图11左上角所示），在一个视图中对比。四个表面的空间对齐帮助Alabi等人确定橙色表面是不同于其他三个。Ferstl等人[15]通过将曲线叠加在一起的方式，比较不同时间段中曲线的集合和它们的集合信息，正如图15b所示。首先使用轮廓变化图[4]可视化在每个时间段的曲线集合（即2D等值线）。然后将不同时间段的图堆叠在一起，从而显示曲线随时间推移的聚类演化。叠加对比有助于跟踪不同时间段的曲线集合。

显示编码使用特定的指标来量化需要对比的对象的不同，并且明确将这些指标的数值加进可视化中。例如，在Wang等人[25]的气象集合可视化作品中，他们计算了模拟场所有网格点上的模拟降水值（来自一个成员）与观测到的降水值（来自卫星）之间的差异，并且将这些不同的值映射到不同的颜色，从而评估集合成员的质量。Köthur等人[30] 扩展了窗口互相关（WCC）矩阵的使用来聚集数据，从而揭露两个时变集合的相关性不同。按照惯例，WCC矩阵的每个单元都存储来自两个不同时间序列（可以是两个随时间变化的集合成员）的两个加窗时间段之间的相关性。在Köthur等人的作品中，WCC矩阵的每个单元格都存储了一组相关值，这些相关值是两个不同时变集合的成员的所有可能组合的相关性（对于指定的时间段）。Köthur等人然后为单元设计图形，从而可视化统计信息摘要（例如平均值，标准差和分位数信息）。

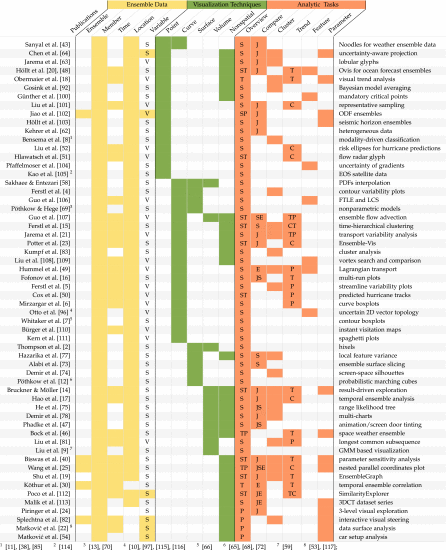
**4.3 聚类**

聚类通过将有相似特征的数据例子放在相同的组中的方法将数据进行分割，这个作为一个有效的方法来揭露突出的数据模式。常用的聚类算法包括：K-均值[78],分层聚类[25], DBSCAN[81]等。对于集合数据，聚类可以用来聚类成员或者对象集合（例如不同位置的值集合）。

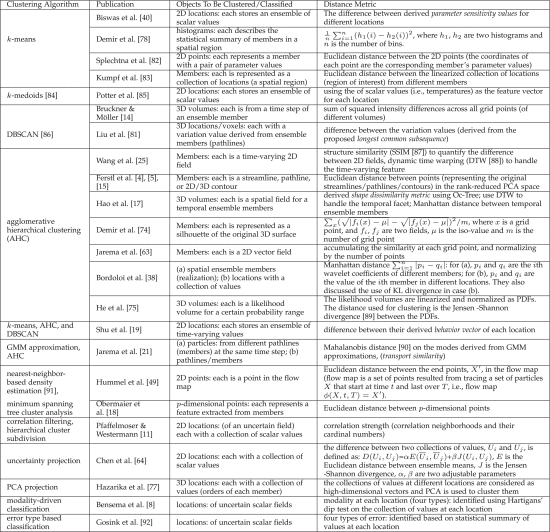
对集合数据执行这个任务的不同是在聚类算法中使用新的距离度量。例如，为了聚类确定性模拟场（每个网格点贮存一个值）中网格点，距离度量可以简单地是存储在这些网格点处的值之间的绝对差。然而，为了聚类一个不确定性场（每个网格点存储一组值）的网格点，必须使用可以测量值集合之间区别的新度量。例如，我们已经看见集合可视化工作，该工作每个网格点的值集合建模成PDF，并使用KL散度来量化不同位置的PDF之间的差异[33]。

从集合可视化文献中，我们发现使用不同集合可视化作品中距离度量是非常不同的。这些度量是非常特定于应用程序的，并且数据特征对其影响很大。因为聚类相关集合可视化作品的数量不是很多，我们决定在本次调查中全部展示。表2中列出了有关这些作品中使用的聚类算法，聚类对象和应用距离度量的详细信息。

**表2.**在集合可视化作品中使用的聚类（分类）算法和距离度量。



**表3.** 基于我们在第2、3和4节中的讨论来组织集合可视化作品，即集合数据，可视化技术和分析任务。 对于“集合数据”列，填充颜色的单元格表示着相应维度大于1。 “变量”列中的S / V表示标量/矢量变量。 “概述”列中的S / T / P表示着空间/时间/参数概述。 “比较”列中的J / S / E表示通过并置/叠加/显式编码的比较。 “趋势”列中的T / C / P表示时间序列图/视图合成/基于路径的方法。 该表列表的脚注属于同一类别。



分类是将数据例子分类成预定义的类别[8],[92],它和聚类是非常相似（即根据它们的相似度将数据例子聚类成不同组的操作）。因此我们在本篇论文中将它们和聚类任务同等对待。对于集合数据，分类的不同在于生成预定义类型的方式。一个典型分类例子（集合可视化作品）是Bensema等人的作品[8](即模型驱动分类)。他们将不确定字段（集合数据中的）的每个网格点的值集合建模成PDF。然后依赖相对应PDF的样式将不确定字段的网格点分成四个类：O类（较低方差分布的网格点），U类（单峰值分布的网格点），B类（双峰值分布的网格点），M类（多峰值分布的网格点）。分类结果揭露了在不确定字段中不同位置中基础集合成员编码的特征复杂度（即模式）。在表2中总结了集合相关分类的作品。

**4.4时间趋势分析**

时间趋势分析任务是为了捕捉感兴趣的数据是如何随时间演变的。在许多领域它是非常重要的任务，像天气预测[25],[40], 流体动力学模拟[18]和物理学中的物质跃迁[17]。对于集合数据，时间趋势可以从一个单独的成员（和传统时间趋势分析类似）或者成员集合（由于集合数据产生的新部分）中提取。根据我们对文献的研究，可以通过三组方法进行时间趋势的抓捕：（1）视图组成；（2）时间序列图及其变化；（3）基于路径的可视化。

视图组成。为了揭露时间趋势，第一组可视化方法依赖于视图组成，即并置或叠加，为了比较不同时间段的集合数据。并置方法通过并排放置对比不同时间段的可视化结果的方式，演示集合数据的时间信息。例如，在Wang等人的气候集合可视化作品中[25]，他们在20个小视图中演示了30个每日降水量，领域科学家可以从中来回看出了解感兴趣区域内降水的时间变化。动画可以被认为是并置方法，因为它也依赖于用户的短期记忆来建立时间段之间的联系。不同在于动画是随时间维度并置视图，而非空间维度。例如，Hao等人[17]在他们的集合可视化中应用动画来展示由高能量粒子构成的3D空间的形状变化。叠加方法为了直接对比，将不同时间段的集合可视化叠加在一起。虽然他们可以容易地导致视图混乱，叠加方法能使用户集中在一个视图，并直接随时间直接对比可视化结果（需要较少的心理定位）。例如，Ferstl等人[15]堆积不同时间段的轮廓变化图（2D可视化结果），从而创建一个时空立方体（图15b），在其中可以轻松感知等值线随时间的演变概况。通常，视图组成方法使用户能够对每个时间段检测足够的详细情况。然而，由于空间的限制或视觉混乱的问题，它们通常不能很好地缩放。

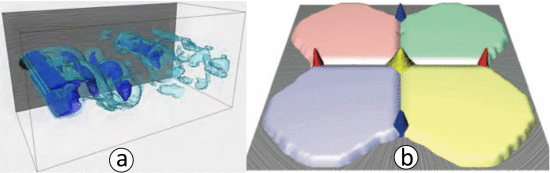
时间序列图。时间序列（时间线）图[20],[46]能有效地总结集合数据的时间趋势和把他们作为高级抽象进行演示。这种类型的2D图使用一个坐标轴来表示时间扩展，另一个坐标轴来演示集合数据中获得的统计测试量[19],[46]。不同的集合成员可以作为2D空间中重叠曲线来表示，所以可以随时进行成员的直接对比。通过捆绑同集群的曲线为条子的方式[15],[21]，也可以完成展示不同时间段中成员的集群信息。可视化结果和并行集[93]相似，可以有效地演示随时间变化的集群的收敛/发散。图15c展示和图15b相同的集合数据的这样的图。从这图可以清楚看出，成员正在演变成四个集群。图15a也是针对集合数据的时间序列图的例子。水平坐标轴是时间维度，两个位置的海平面的值分布沿着垂直坐标轴展示。值的两个集合（两个位置的所有集合成员）作为两个并排的PDF进行建模。通过随时间对比和跟踪这对PDF，可以观察到左右两侧PDF代表的两个位置的风险水平的时间趋势是不同的。左侧通常有低风险水平，而右侧有增长风险水平，它从绿色（低风险）变化到红色（高风险）。时间序列图通常覆盖长时间范围，但是他们展示的时间趋势通常是高级抽象。

路径演示。对于随时间变化的矢量场集合数据，路径是直观的方式来展示不确定性的基础的时间趋势。例如，Cox等人[50]使用一系列路径来展示预测的飓风轨迹。在他们的作品中，一个路径代表着飓风的一种可能轨迹（一个集合成员）。将不同成员的路径重叠起来会展示飓风将影响的高/低可能性区域。当成员维度很大时，基于路径的方法可以很容易导致视图混乱。因此，不是直接演示路径，而是在某些情况将显示这些路径的摘要统计信息。不确定度圆锥[42]和曲线箱线图（应用于路径线时）[6]是例子。基于路径方法可以直观演示时间趋势和不确定性演变，这是因为它们可以直接编码进模拟场。然而，由于缺乏路径的同步（在不同时间段），它们只能受限在矢量场集合中，并且不能展示特定时间段的集合成员的详细信息。

**4.5特征提取**

特征提取为了从模拟场中提取某些感兴趣的几何/拓扑特征。和我们在第三节讨论的四个基本几何特征不同，本节的特征会更复杂，例如流动行为（例如涡流，涡流），拓扑临界点（如源、汇、鞍）等。

对于集合数据，这些特征需要从不确定性字段中提取。他们的可视化不仅要揭露特征本身，还包括关联的不确定性。这里，我们使用局部涡旋检测的例子来阐述特征提取是怎么样扩展到集合数据的。对于一个确定性的矢量场，局部涡旋检测器对每个网格点计算一个数值准则（例如Q准则[94]，lambda \_2准则[95]等），并且用比较用户指定阈值的标准值来确定网格点是否属于涡旋区域。然而，不能直接从集合数据的不确定字段中计算出数值尺度。Otto等人[10]通过使用蒙特卡洛方法解决了这个问题。考虑到相近网格点的空间相关性，他们首先每个网格点的集合数据建模成高斯分布。然后从分布中获得蒙特卡洛样本，并且计算每个样本的数值尺度。结果，每个网格点都会有标准值集合。将这值集合和用户指定的阈值比较,他们获得了网格点的概率值，表明了网格点在涡旋区域中的的可能性。对所有网格点进行操作，他们为原始不确定矢量场派生出涡旋概率场，该场可以通过传统的等值面或体积可视化技术进行可视化。图16a展示了具有lambda \_2的涡旋概率场的等值面可视化，其中P（λ2<-0.003）。



**图16.**

（a）圆柱体集合数据集（时间步长 100）：等值线0.05和0.95的概率场P（λ2<-0.003）的等值面可视化。 图片由Otto等人提供。 [10]©2012 John Wiley＆Sons，Inc.（b）vc的LIC和不确定的临界点：红色/蓝色/黄色的峰是源/汇/鞍点。 图片由Otto等人提供。 [96]©2010 John Wiley＆Sons，Inc.。

确定关键点（即源，汇，鞍）是分析空间场的拓扑结构的有效方法。然而，和关键点有确定性位置的确定性字段不同的是，集合数据的不确定性字段给关键点的提取带来挑战。针对这个问题，Otto等人[96],[97]将流线的概念归类到不确定性矢量场，并且提出了确定字段的关键点的算法。在他们的作品中，他们将不确定性矢量场建模成密度分布函数（即每个网格点的矢量集合被建模成高斯分布），并且对功能进行流线跟踪。按照惯例，对于特定的集合步骤，流线跟踪算法会得出粒子的确定性位置。对于不确定性场，确定性位置变成了不确定性位置，并且这些位置在每一个整合步骤中构成粒子密度分布。基于不同整合步骤中的密度分布的导数，Otto等人从不确定性场中提取了源，汇和鞍点。他们还执行了蒙特卡洛方法来整合概率粒子路径，以便他们可以分割不确定性场。正如图16b所示，不确定性源点，汇点和鞍点以红色，蓝色和黄色的（高度场的）峰值表示。不同颜色的区域展示了分割区域（受不同来源/汇点影响）。

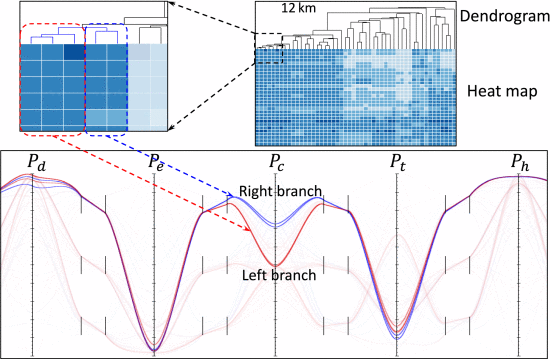
**4.6参数分析**

参数分析任务是为了建立参数数据和集合数据的联系，从而分析模拟输入的不确定性。从领域科学家的角度看，参数数据和集合数据（模拟输入，输出）是紧密关联，并且为了建立双方的联系应该在两边进行分析。

Sedlmair等人[98]为参数分析提出了概念设计框架。和他们不同，我们在本次调查只关注和集合数据相关的参数分析作品。和集合数据没有关系的纯参数分析作品已经被排除。集合数据的可视化已经在第三节彻底地讨论了，因此在这节我们关注讨论参数数据是怎么可视化的，和这两类数据是怎么关联的。

可视化参数数据。将单个参数视为维度，有不同值的多个参数构成了高维空间。因此，参数可视化可以采用高维数据可视化技术。根据我们的报告，参数可视化最常用的方法包括平行坐标图（PCP）[18]，[24]，[25]，[44]，轮图[99]，散点图[22]，[82] ，折线图[40]，[82]和图形[46]。大多数方法不仅展示了参数空间，还演示了不同参数之间的相互关系。例如，PCP可以用一个平行轴表示一个参数，带有折线的参数设置（一次运行）横跨所有轴。然后PCP邻近轴形成的关系形式可以演示对应参数之间的关联性。

连接集合数据。每个参数数据例子（一个参数设置）和每个集合成员是一一对应的。领域科学家通常通过这种对应关系研究模拟参数，从而连接参数数据和集合数据。通常通过协调多个视图来同时展示参数数据和集合数据。具体来说，刷取和链接是连接参数视图和集成视图的最常用策略[40]，[62]。另外，我们注意到，在关联参数数据和集合数据时,发现聚类之间的关系（即参数设置的聚类，集合成员的聚类）[14],[18]。在图17中，参数数据在PCP的变体中展示，即嵌套的PCP [25]；集合数据通过树状图（每个叶子是一个集合成员）和热图（水平/垂直扩展代表成员/时间维度）的链接可视化进行演示。通过链接可视化，领域科学家可以注意到P\_c数值（一个输入参数）是怎么样影响模拟输出。



**图17.**

在参数数据和集合数据之间建立连接。 热图的每一列代表一个时间集合成员，并且用层次聚类对成员进行聚类。 两个突出显示的成员组（在蓝色和红色框中）之间的差异是由参数Pc值的差异导致的。 图片由Wang等人提供。 [25]©2017 IEEE。

**第5节**

**讨论**

在上面两节中，我们将可视化技术分成五类，并且总结了六个集合可视化的分析任务。从分类和总结中我们发现一些研究趋势和一些有希望的研究方向。这些将会在这节和我们的作品的局限性一起讨论。

**5.1观察，研究趋势**

不确定性建模。许多可视化作品通过将集合成员建模成PDF的方式分析集合数据的不确定性。我们发现许多可视化研究在提高PDF的准确性来反映不确定性付出了许多努力。这些努力有两个方向：（1）允许更多的方式；（2）考虑PDF之间的相关性。关于第一个方向，广泛使用在集合可视化的高斯分布用于建模成PDF。然而，基础的集合成员可能不遵循正态分布[13]。为了更好地建模不确定性，研究人员开始寻找有更多形式的更复杂的模型[8],[63]。这个方向也改善了非参数模型（例如核密度估计，直方图[2]）和半参数模型（例如GMM [9]，[21]）的使用。对应第二个方向，研究人员已经开始考虑邻近网格点的相关性[10],[116]。科学数据集通常使用局部数据连续性的固有属性来测量连续的物理现象。结果，假设各个网格位置处的PDF是独立的，可能会导致不确定性的错误建模。Pöthkow等人[12]和Pfaffelmoser等人[72]已经研究这个错误表示，并且得出结论，即考虑不同网格位置的PDF之间的空间依赖性是很重要的。因此，后一批基于分布的集成可视化工作更集中在多变量分布上，以便于更好地建模不确定性[12]，[13]，[69]，[72]。

协调多个视图。我们也观察到在集合可视化作品中协调多个视图比单个静态视图更经常使用。事实上，几乎不可能在单个静态视图中演示集合的所有方面。这些尝试将主要导致信息过载和/或视觉混乱。相反，不同视图有不同的关注点，协调多个视图可以演示集合数据的不同方面。然后，分析集合数据将要求在这些视图之间进行探索的过程，并且驱动该探索的交互作用（例如，笔刷和链接）。可视化正在从单个静态视图演变为交互式视觉分析系统的这种观察/趋势不仅限于集合可视化作品，还包括可视化作品的其他类型。

**5.2未来的挑战和机遇**

集合可视化依然年轻[35],[57],并且针对它的发展可以进行许多工作。我们讨论一些我们相信重要和有希望的方向。

可变和集合维度。从我们结构化文献分析中，我们发现集合数据的多元特征还没有充分地探索。尽管有些作品可视化集合数据的多个变量[22], [82], [117],但是分析这些变量的相关性和关联不是他们的关注点。鉴于变量相关性分析是传统科学数据的如此重要的主题，我们相信在集合数据的背景中探索它也是很有希望，该背景下变量相关性将会变成不确定性相关性。另外，许多现存的可视化作品也没有覆盖集合数据的集合维度。关注这个维度可能导致新的研究思路和开拓性工作。虽然它也依赖这类数据的可用性和领域科学家的特定需求。

专家评估和技术可移植性。我们也相信在集合可视化技术的评估中会有更多改善被执行。一个新技术的有效性通常由领域科学家评估。然而，在进行严格评估中，仍有一些严重问题存在。例如，有多少领域科学家应该参加评估来消除主观影响？领域科学家是多愿意应用一个新技术？熟悉一个新技术需要多少时间？对这些问题的回答或者对评估应包括哪些方面的系统化将大大提高评估的严格性。另外，一个提出的集合可视化技术是怎么容易应用到其他领域的相似问题，即技术的可移植性，在绝大多数情况不能很好地被讨论。解决此问题将显著地改善集合可视化技术的可重用性。

可扩展性。集合可视化的可扩展性问题可能一直存在，因为模拟现实世界物理现象的尝试始终旨在获得更高的空间和/或时间决议。领域科学家目的是生成更多集合/成员来更好地获得他们模型的基本不确定性，这将会显著增大集合数据的规模。因此，更简洁的数据表示，有序的记忆结构，高性能的数据查询，智能的数据移动策略，这些需要来支持可视化的增长规模。从我们的报告中，我们看到许多通过并行性来提高集合可视化的表现的作品，例如[81], [107]。可以基于特定的分析目标跨集合数据的不同维度进行并行化（例如，跨成员，空间位置或时间段并行化计算）。在数据存储方面，大多数现有的基于分布的基于数据减少方法是针对集合数据，例如 [118]，[119]，他们可以直接在集合数据中使用。然而，使用这些方法通常是基于基础成员分布的特定假设，并且可能需要采用某些修改才能更好地保留集合数据的特征。此外，我们相信对于集合数据的现场可视化也是个解决可扩展性挑战的有希望的方向[120]。

**5.3局限性**

作为文献的结构化分析，我们的受到了和这类作品相关的标准限制，即他们的主观性。我们用于查看集合可视化文件的视角，用于选择它们的标准以及用于对它们进行分类的分类法均受我们了解集合可视化方法的影响。这些主观的选择和决定不是最好的或唯一的。毫无疑问的是存在着等效的分析结构，不同角度的分类法。然而，我们也相信我们的作品是足够为当前集合可视化的研究提供有用的图片。鉴于只有很少集合可视化的研究作品，我们希望我们的工作能扮演满足严苛需求的起点，并且在未来导致更全面的总结或者该主题的系统化。

在有限空间中，我们更关注总结不同集合可视化技术的共同特征，和证明我们提出的分类法的有效性。然而，我们相信我们的研究可以通过对比相同类型或不同类型的可视化的方法来进一步扩展，例如讨论在不同应用场景中不同技术的优点和缺点。这将潜在地导致集合可视化的概念设计框架，并且助长更多该主题的理论分析。

**第6节**

**结论**

本论文展示了集合可视化的调查。该调查将集合可视化技术分成五类并且回顾了传统可视化技术是怎么样应用到集合数据中的。它也阐述了由于集合数据的独特特征，六个分析任务是怎么样不同地执行的。通过本次调查，我们将代表不同可视化作品的点连接在一起，并且为这个主题树立了一个大格局。我们希望我们的作品可以帮助到集合可视化的理论分析和支持新型集合可视化技术的发展。