时空中的无监督学习

基于图形技术和深度神经网络的现代计算机视觉方法

第一章 无监督视觉学习：从像素到视觉

前言

在第一章中，我们介绍了无监督学习的七个原则，然后简要介绍了与这些基本原则和概念密切相关的之后几章所涵盖的主题。

目录

正文

1.1 “看见（see）”意味着什么?

试图想象我第一次睁开眼睛的时候世界是什么样子的。我看到了什么？我很确定我非常清楚地看到了不同色彩的每个像素。然而，我看到的是物体还是人？当我看着妈妈的时候，我看到了什么？当然，那种温暖和明亮让我感到亲近，让我很有安全感，但我对妈妈了解多少呢？我能看到她美丽深邃的眼睛，或者她长长的黑发和最迷人的微笑吗？我担心我可能错过了所有的这些，因为那时我真的不知道什么是“眼睛”、“鼻子”、“嘴”和“头发”。我怎么能在甚至不知道我是什么、或者人类是什么的时候，把她看作一个人，一个和我一样的人？

因为我不能把任何事情同过去的经历联系起来，所以没有什么事情是我能“认出”或看到的。这世上没有“能看到的东西”，因为图像的不同部分之间还没有建立起联系。像素仅仅是像素，对其他东西而言，我大概都是瞎的。仅仅观察到一些具有相似颜色的像素组并不意味着我可以将它们视为某种东西。任何事物背后都没有能把它和其他事物联系起来的、赋予它们意义和作为事物的“真相”的“过去”。

现在，当我看着我的母亲时，我可以看到她是谁：有很多的经验将我们联系在一起。她在我人生的不同阶段都有很多的形象，也有很多我当时的形象。所有这些记忆都紧密联系在一起，在一个空间和时间都连贯的故事中相互作用。所有这些母亲的形象，通过她在我生命中的独特轨迹而联系在一起，是她在我的世界中成为她自己。这些形象让母亲成为我看着她时所看到的样子。

现在，当我看到母亲脸上的一个像素时，事实上我同时也看到了更多的东西。我看到一个皮肤像素、一张脸和一个人体像素。我也看到了我自己母亲的一个像素——我生命中那个特殊而独特的重要人物。然后，在空间和时间的更高层次上，那个像素也是我们的一部分，我和她，母亲和儿子，这一从一开始就存在的联系。我还看到了一个属于人类、生命和地球的像素，它围绕着太阳运行。直到现在，我才能看到现实世界的所有层面，而这需要很多年才能做到。

我现在的视野比最初的时候更深，最初的时候它几乎不存在。所有这些视觉层包含对象和对象的一部分、交互，以及当前存在、同时也是真实存在的活动。回到我第一次睁开眼睛的时候，世界刚刚开始活动。我一定有一种深深的、强烈的冲动，把我拉向意想不到的灯光，伴随着它们令人惊讶又诱人的闪烁模式。但我当时对接下来会发生什么了解多少呢？我现在看到的一切都是我们的故事。这是我自己作为世界的一部分的视觉故事，因为我越来越深刻地认识这个世界，并希望能够想象到以前会是什么样子、接下来会发生什么。

1.2 什么是无监督学习?

尽管在最重要的时刻，妈妈总是在我身边，教给我一些我需要知道的十分重要的知识，但在我生命的最初几年，她绝对不是教我如何去“看见”的人。她没有把我视野中的每一个像素都赋予众多层次的意义，而这其实是根本不可能的。是我的大脑教会了我如何在学习之后，从我在这个世界的众多经历中以一种自然的、无监督的方式去“看见”。是我的大脑获得了我感知到的每一个像素，并赋予它们意义和价值。是我的大脑复制所有这些像素，并将它们排列在不同的或高或低、同时存在的视觉空间和时间层上，让这一切变得对我来说都是真实的。所有这些层面，不管是现在还是过去，都在一个连贯的故事中找到了共识。正是在这个故事里，我赋予了此时此地一种意义；正是在这个时空世界里，我“看见”了。

从这个基本的角度来看，理解无监督视觉学习将帮助我们理解一件事：为了更好地关心我们的世界、改善我们的生活，我们需要学习更多、看到更多关于这个世界的一切。出于同样的原因，自然世界中的无监督视觉学习也是理解智能、明白大脑如何工作以及我们建立真正智能的系统的基础。这些智能系统将会学会像人类一样去“看”事物，然后学会与人类和谐相处。

无监督视觉学习可能是人工智能研究人员解决的最有趣的问题之一。怎么可能实现去学习这个世界，学习这个世界的所有属性，学习这么多以简单或很复杂的方式相互作用的不同类型的物体？最重要的是，不了解真相怎么可能学会这些？这能做到吗？我们还没有答案，但我们知道的是，孩子可以在父母最少的干预下来学习这个世界。当父母或老师干预教育时，他们实际上什么都没有告诉我们。他们肯定不会用语义标签标记我们视野中的每个像素。从我们生命的最初几个月开始，我们的大脑开始自己学习将像素排列、并分组到不同的区域，这也是很有意义的，因为我们不断地获得经验。至少，在最开始，我们的经验不涉及复杂的物理相互作用，我们的视觉学习主要基于观察。

虽然与世界交互对学习至关重要，但在这本书里，我们感兴趣的主要是通过从时间和空间角度观察世界，我们能从展现在我们眼前的东西中学到什么。虽然我们在最后一章简要讨论了交互的主题，并讨论了我们如何采取行动并从成果中学习，但我们认为，应该首先将注意力集中于从我们无法影响的时空数据中学习的更有限的情况，以便更好地理解无监督学习的基本限制是什么。我们应该做哪些假设？我们需要访问什么类型的数据、需要多少数据？我们可以学习哪些类和概念？什么类型的计算模型可以解决无监督学习问题？

正如我们在整本书中所展示的，这些基本问题可以揭示普遍的答案，这可能成为在计算机视觉和机器人学中实现许多现实世界应用的实用工具。在本书的最后，我们将冒险进入想象力的世界，并设想一个可以在空间和时间中自学的通用计算机视觉系统。从一开始，我们将建立一套无监督学习的一般原则，我们将通过具体的任务、算法和广泛的实验评估来逐章演示这些原则。最后，我们将展示如何使用这些基本原则来构建一个通用系统，该系统可以在一个统一的视觉故事网络中独立学习对时空世界的多层解释。

到本书的最后，我们会更好地理解无监督学习实际上是在自然世界中的学习，它不能在没有来自浩瀚数据海洋的数据输入的情况下自己发生，这遵循物理和经验统计规律。无监督学习不仅仅是快速高效的算法或某个计算机模型的架构，它还与该模型运行的世界密切相关。最后，我们必须使其达到与外界交互的学习水平，而这个学习系统本身也是外界的一部分。从这个角度来看，计算机视觉中的无监督学习成为了了解关于学习、关于我们自己，以及关于我们如何发现和“看见”我们周围的世界的绝佳机会。

1.3 时空中的视觉学习

视野如此丰富，一幅画可以诉说千言万语。视觉也是我们接触世界的第一个窗口，是我们最重要的感觉。视觉发生在空间和时间中，创造出我们看到的一切，从静止或运动的物体到发生在场景中的活动，以及将所有成员和他们复杂的交互联系在一起的整个故事。视觉几乎拥有一切，从最简单的颜色像素和最普通的物体，到最狂野最深刻的想象，视觉试图创造和反映我们生活的世界。作为一个序列，视觉必须考虑物理规则和经验统计规律，这些规律赋予自然世界在空间和时间上的连贯性和一致性。因此，视觉必须建立在反映这种一致性的某些分组属性和统计原则的基础上。如果我们想真正有机会在野外以不受监督的方式学习，我们就必须理解这些原则，并将其用于建立的计算模型。

在空间和时间上思考具有一定的优势。我们周围存在和发生的一切都在时空中，没有什么是真正静止的。然而在今天，计算机视觉研究在很大程度上被专注于单个图像处理和识别的方法所主导。很少有直接从视频开始的方法。这更多是出于历史原因：处理单个图像的成本更低，并且人类表明在单个图像中进行识别是可能的。所以，如果人类做到了，而且成本更低，为什么不让程序做同样的事呢？我们认为单图像任务在有监督的环境下更有意义。如果要在无监督的情况下学习，那么我们更应该考虑现实世界中的事物是什么样的：物体与更高层次的实体，动作与交互，复杂的活动与完整的故事，都是在时空中存在的，而时间与空间有着深刻的联系。每个物体从一个时刻到下一个时刻，在外观或位置上都有一点点变化。在每一个地方，都有一种变化的因素，一种时间和空间上的振动，而我们应该从中学习。物体的运动通常与其周围环境不同。它们看起来也和它们的背景不同。因此，变化从一开始就在时间和空间两个维度上共存。同时，物体在空间和时间上都是一致且连贯的。它们的运动通常变化平稳，它们的相互作用遵循一定的模式。它们的外观也平滑地变化，并遵循某些对称的几何和颜色模式。因此，在我们的研究开始时，考虑视频作为输入似乎是必须的。

只有当我们考虑时空现实时，物理世界的所有这些属性才能被利用。如果我们想从根本上解决无监督学习的难题，就必须充分利用自然世界中普遍适用的每一条信息和属性。

将视频作为输入，从一开始就同时考虑空间和时间，还有一个实用的优势。我们可以免费获得大量未标记的视频数据，能够自动标记这些视频的使用无监督学习的任何解决方案，相比起需要非常昂贵的手动注释的严格监督的方案都具有巨大的优势。

因此，进行无监督学习的能力对学术研究和工业发展都非常有用。此外， 与单个图像相比，可用的视频数据量的增加有可能极大地提高泛化能力，并允许从一开始就同时学习对象及其交互。对象类通常由它们在更大的故事中的行为和角色来定义。一个物体在它的局部层次上的属性和它在全球时空环境中所扮演的角色之间应该具有明确的一致性，如果我们从一开始就考虑时空域，我们就只能利用这种一致性。

1.3.1 当前无监督学习的趋势

在机器学习、计算机视觉和机器人研究中，人们对无监督学习的兴趣正在稳步增加。经典的研究结果基于这样的观察——真实世界的数据基于特定的核心、固有的属性，如颜色、纹理、形状，自然地分组到特定的类别中。因此，基于这些属性的相似元素应该属于同一个组或集群，而不相似的元素应该放在不同的集群中。因此，在过去的五十年中，Gan等人提出了大量的算法，机器学习中非常广阔的聚类研究领域从此诞生，这些领域可分为几大类：（1）与K-means相关的算法Lloyd、EM算法Dempster、有一个明确的概率公式，并试图最大化数据的可能性条件下的课堂作业的算法；（2）直接优化聚类密度的方法，如均值漂移算法Comaniciu and Meer，Fukunage and Hostetler，以及基于密度的空间聚类（DBSCAN）算法Ester；（3）在贪婪的聚类中，从较小的子聚类中形成聚类的层次方法fashion Day和Edelsbrunner，Ward Jr，Sibson，Johnson或分裂聚类算法（DIANA）Kaufman和Rousseeuw，它们从大的聚类开始，迭代地将较大的聚类分成较小的聚类；（4）谱聚类算法，它是基于邻接矩阵的特征向量和特征值，或与数据点Cheeger、Donath and Hoffman、Meila and Shi，Shi and Malik，Ng等相关的图的拉普拉斯算法。在本书中讨论并应用于不同的计算机视觉问题的聚类算法大多与谱聚类方法有关。

直到不久前，出于充分的理由，大多数无监督学习研究都集中在提出和研究各种聚类算法上，Duda等人对此进行了研究。大多数无监督的学习任务都隐含或明确地需要某种聚类。所有研究机器学习的人都希望，即使没有机器学习，世界的整个结构，包括它被分成特定的类别的以不同的方式移动、联系和行动的实体，应该在一个纯无监督的学习环境中自然出现。这种具有良好实体和关系的结构的发现意味着某种数据聚类。此外，有洞察力的读者肯定会在本书中发现，这里提出的方法的核心也是基于聚类原则。

与20年前更普遍的聚类方法相比，目前在无监督学习方面的研究更加通用、多样和复杂。然而，无监督学习仍处于起步阶段，在无监督学习这一神秘未解的拼图中还缺少许多部分。还有很多没有答案的问题，甚至还有一些未被提出的问题。在这一点上，我们开始意识到时空域提供了比单个图像更大的优势，因为时间维度在聚类时带来了重要的信息。不同的物体不仅外观不同，运动方式也不同。属于一体的物体不仅看起来相似，移动方式也相同，而不属于一体的物体则在时间和空间上分离。使世界结构具有额外一致性和连贯性的时间维度，突然成为无监督学习难题中的关键角色。

因此，毫不奇怪，最初的无监督学习的计算机视觉专用现代技术专注于时空、视频领域。例如，在一篇开创性的经典论文Sivic和 Zisserman中，作者提出了一种检索视频中的对象的算法，该算法基于匹配关键点来发现视频中的给定对象，这些关键点在后续帧中的外观和几何形状是稳定的。这些稳定的关键点构成的集群与单个物理对象相关联。虽然这篇论文不是专门研究无监督学习和聚类的，但它实际上在很大程度上依赖于无监督学习和聚类，来完成从视频中检索对象的任务。在我们早期关于视频中物体发现的工作中，Leordeanu等人采用了和我们类似的方法，通过匹配视频帧的关键点簇来发现对象，这些关键点簇在足够长的时间内保持几何稳定。我们在实验中注意到一个有趣的事实：当一组关键点在特定的时间内保持几何稳定时，它们确实属于单个对象的概率会突然从近乎于 0 增加到近乎于1——这再次表明，在无监督学习游戏中，时间可以提供很有效的线索，告诉我们什么应该在一起、什么不应该在一起。

自从第一批以无监督方式发现视频中对象的方法出现以来，其他研究人员也开始研究这一方向。在视频中发现对象的任务越来越重要，如今，大多数方法都是在深度学习的背景下制定的。似乎有几个研究方向与使用无监督的方式从视频中学习对象有关。

1.3.2 与格式塔心理学的关系

当前许多无监督学习的方法具有很强的关联性，有些甚至受到20世纪初奥地利和德国建立的格式塔心理学学派的启发。格式塔心理学中的“格式塔”是德语单词，意思是“元素或模式的配置”。这一学派介绍并研究了这样一个主要观点，即物体是从部分中产生的整体，而不是部分的简单加和。这就是“整体大于部分之和”这句话的由来。因此，格式塔认知科学家提出并研究了集中“分组法则”，大脑则用它们来形成这样的“整体”。这些分组法则包括，例如，描述接近的元素应该属于一个整体的概率法则，或者描述相似事物应该被分组在一起的相似法则。同样地，表现出对称、几何连续性或或者有相似的运动趋势的元素也可能属于一个整体。除了这些将更小的事物组合成为更大的事物的原则之外，格式塔心理学还研究了我们有意识地将事物视为一个整体、并根据已有的经验来解释它们的方式。

下面我们将展示关于无监督学习的关键成果，我们将其归纳为一系列原则，这些原则在统计学意义上被认为是正确的，但它们不一定总是正确的。这些原则与最初的格式塔定律密切相关，在某种意义上，它们可以被视为在现代机器学习和计算机视觉的背景下对这些定律的重新解释。虽然格式塔原则主要体现在分组的初始阶段，但我们更进一步，从计算得角度提出这些原则，以便最终建立学会自己“看见”的人工系统。

1.4 无监督学习的原则

原则1：物体是全局场景中的局部离群点，体积小，外观和运动与它们的大背景不同。

物体通常看起来比包含它们的场景小，这是有一定的道理的。通常情况下，物体有着不同的外表，与它们直接的背景形成对比。让我们想想一朵花在绿色的田野上，它有着突出的颜色，又或者一颗红色的水果在树上等着被摘下来。甚至天空中的太阳也有很强的黄色光芒，与更大的蓝天形成对比。诚然，人们眼中物体的颜色是我们视觉系统的产物。然而并非偶然的是，大脑的视觉部分创造了对颜色的感知，使这些物体与周围的环境形成对比。视觉可以发现物体、了解物体，然后就更容易识别它们。当然，这种统计观察也有例外。拥有惊人的伪装能力的最著名的动物——变色龙，就是一个例子。但是如果所有的物体都被伪装起来，智能视觉和物体识别技术将很难发展。

物体相对于场景也有不同的运动。它们通常比静态背景更灵活，移动方式更快更复杂。事实上，对于生物来说，生物越小，行动就越快、越复杂。较小的鸟比较大的鸟有更快的改变方向的能力。生物视觉系统被赋予探测运动的能力并非偶然。光流的估计，即图像中像素的视运动，是计算机视觉中的一个基本问题，也是任何机器人视觉系统的基本能力。对于每一项涉及运动的任务，几乎都要使用光流。

原则2：通常可以用高精度的算法（不一定有高召回率）来挑选出属于单个对象或类别的数据样本。

事实上，很多时候我们可以发现属于一个单一对象或实体的场景部分或数据流。虽然对象类别或实体类型可能仍然未知，但我们几乎可以确定样本（可以是像素、面片、整个区域或其他数据样本组）确实属于同一类，这是基于某些分组属性得出的，这些属性使得同一个分组内的样本很可能是相同的。

让我们在研究一种我们第一次看到的新动物时想想这个例子。当我们为了理解动物的独特特征而了解它的新特性时，我们会首先应用已有的知识。我们可能会先从动物的形象中分出一些我们能够识别的部分。我们可能会认出腿在哪里，认出皮毛、头和身体的其他部分。最后将这些部分放在一起，我们就第一次认识了这种动物的整体外观。类似地，通过将它的行为模式分解成可识别的片段，我们可能最终就能认识它独特的行为模式。如果可能的话，根据它的外貌、身体特征和行为，我们以后可能会决定它属于哪个大类。虽然起初我们对这种新动物一无所知，但通过识别出几个同时聚集在一起的小部分，我们有了这样一种想法：在那个空间区域里有一只动物——一只单一的动物。让我们想一想：当它们都来自不同的动物时，在空间中看到非常接近的腿、身体和头被一些皮毛聚集在一起的可能性有多大？可以肯定，所有这些部分构成了一个单一的整体，而它独特的外观可以让我们意识到，我们似乎看到了一个新的物种。

在本书中，我们将看到许多像上面的例子一样的情况，几个已知类在同一空间和时间区域中的一致共现是新的类存在的指示，并且可以可靠地以无监督的方式用作训练新分类器的积极情况。

原则3：物体在空间和时间上显示出对称、相关和一致的特性。利用基于外观、运动和行为的分组线索，可以获得高精度的正目标样本。这样的线索，很可能属于一个单一的物体或类别，被称为高度可能的积极特征（HPP）。

我们人类知觉中用来在视野中形成物体的分组线索，实际上极有可能是高概率正特征（HPP）——它们极有可能将属于同一场景的部分聚集在一起。场景中这些部分的像素具有相似的颜色分布、相似的纹理和对称的形状。这种在颜色、形状和外观上的一致不太可能是偶然发生的。这样一组看起来相似的像素开始一起平稳移动，同时改变方向，并在背景场景中清晰地突出出来，这也是非常不可能的。分组在时间和空间上可能不会给我们每一条对一个对象的信息，但是它可以给我们足够又高精度的（不一定高召回）的积极训练情况来开始学习过程，此时对背景的消极训练案例通常是很容易收集的。

原则4：物体在它们的时空近邻形成明显的运动轨迹和外观模式。

这个原则是紧接着前几个原则而来的，它似乎涉及很多直观的常识。一个物体应该是在空间和时间上一致和连贯的东西。如果我们给空间增加第四个维度——时间，那么这个4D世界内的物体的物理部分将会非常紧密地联系在一起。因此，在视频帧的2D（图像空间）+1D（时间）世界中，我们期望在图像序列中同样紧密的联系，因为每一帧都是3D世界中的投影。我们应该能够将物体描述为空间和时间上的点簇，这些点簇通过相似的运动轨迹和外观模式而紧密相连。很明显，这些集群应该相对于给定的规模和时间的邻域来看。比如，一辆汽车形成一个很强的集群。但是整个地球以及所有人类、动物和物体也是如此，但是规模会大得多。而地球也是太阳系中的一个物体。

原则5：偶然的对齐很少见。当它们发生时，通常表示模型和图像之间的正确对齐。对齐可以是几何的，也可以是基于外观的，虽然很少见，但当它们发生时，会形成一个强大的集群，以多种方式相互加强。

这个原则是无监督学习的核心。对齐，我们指的是复杂的对称性、形状的精细几何对齐、在运动模式中许多独立分类器输出或聚合的共现，以及在给定的局部时空邻域中的出现。对齐可能是偶然发生的，但并不总会发生。对齐在时空中非常罕见，除非存在一个产生导致对齐的实体。这就是对齐是应该用来开始学习新的物体和其他视觉类别的HPP特征的原因。

原则6：当几个弱的独立分类器一致地一起触发时，这就表明存在一个更高层次的类，可以为这个类学习一个更强的分类器。

与前面的原则紧密相关的是，许多独立的分类器不太可能紧密地结合在一起，除非它们在语义或抽象的更高级别上有相同的东西（可能是未知的类）。例如，不同的分类器可以在对象的不同部分触发，而不存在一个可以“看到”整个对象的分类器。有的分类器可以在腿部触发，有的分类器可以在鼻子触发，还有的分类器可以在头部或皮毛触发，但不能面对整只动物触发。当这种情况发生时，一个更强的分类器可以通过利用分类器组的共射来学习观察整只动物。

原则7：为了改进或学习新的类，我们需要增加训练数据的数量和难度，以及增加分类器的功率和多样性。通过这种方式，我们可以通过使用现有分类器之间的聚合来作为对新分类器的教师监督信号，以无监督的方式进行几代分类器的学习。

新的类可能是比我们已经知道的类更高语义级别的概念，也可能只是我们不知道的不同类别。无论哪种类型，为了学习一个新的类，我们只能使用已有的分类器，因此我们将依靠这些分类器一致的（不太可能）协同发射来发现时空区域，在那里我们可能捕捉到一个新物体或类的存在，也就是说，到目前为止我们还看不到。因为我们还没有学会如何来“看见”它。学习一个新的类需要现有分类器的协同工作，但也需要足够的数据来支持新类的分类。因此，必须添加新的更有挑战性的数据（包含难以发现的案例或新类别）。此外，随着输入的未标记训练数据变得越来越具有挑战性，一个重要的任务是要提升分类器的能力和多样性水平。因此，我们可以用几代旧的分类器协同工作，来为新的但更强的分类器提供监督。一代又一代的分类器可以进化，从智能识别非常简单和局部的对象，发展到对周围的视觉世界形成一种近乎完整、圆润、连贯的感觉——成为一个普遍的无监督学习系统。

原则4的应用：视频中目标分割的无监督学习

到目前为止，我们已经建立了基本的元素、思想和算法，为在空间和时间上实现无监督学习迈出了基本的一步。当视频中的新物体在空间中运动时，在没有人类监督的情况下发现这些新物体是视觉学习中最有趣和最具挑战性的问题之一。仅仅通过观看视频，我们能学到多少东西？当有大量的视频数据可用，但它们没有任何正确的标签时，我们只是被动的观察者，那么我们学习的极限是什么？这个例子似乎是最难的，但为了更好地理解在时空领域中学习的局限性，我们选择了研究它。我们在本节中介绍的工作，其核心是基于我们在开始时介绍过的无监督视觉学习的关键原则之一（原则4）所捕获的观察：

原则4：物体在它们的时空近邻形成明显的运动轨迹和外观模式。

尽管人们倾向于在描述空间和时间上的某个特定区域的视频拍摄中，对哪个是主要物体这一问题达成一致，但我们尚不清楚这是如何做到的。有许多有趣的问题等待我们回答：哪些特征使一组像素在给定序列中脱颖而出，成为一个单一的主要对象？是点轨迹的模式、共同的外观、还是与背景的对比，使一个对象突出作为一个单一的实体？还是所有这些因素的结合？什么是最好的结合方式？为了开始学习，我们是否需要广泛的预训练或一些固有的、更高层次的"客观"特征，或者原则上我们可以从零开始？

我们观察到世界在空间和时间上是一致和连贯的，被很好地构造成具有许多分组特性的物体，早期心理学家也认识到了这一点。既然视频是对四维世界在时间和空间上的投影，那么这些投影应该继承思维世界的许多特性，同样地，大脑也应该做好充分的准备来利用这些投影。接下来，我们简要介绍了在空间和时间中发现物体的最初方法，这一方法通过广泛的实验验证了：使用非常简单的运动和外观特征有可能发现物体在空间和时间中的集群。该方法足够通用，并且可以自然地扩展，以包含更多预训练的特性。

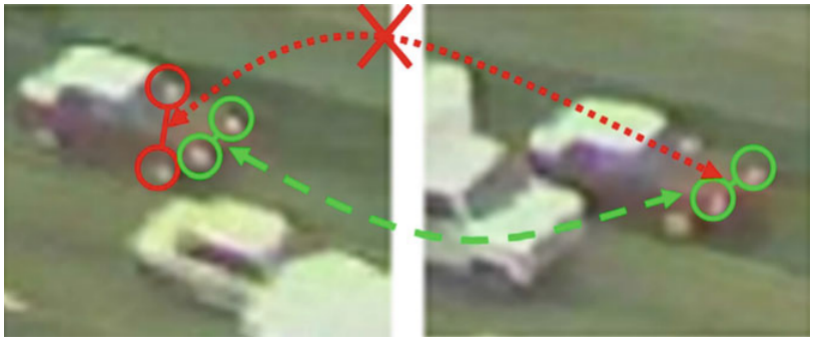
我们的方法基于物体的双重视角，以互补的方式将空间和时间维度相结合：（1）一个物体的同一点，在不同的时间点被不同的像素捕获，再通过长距离的运动链连接这些像素点——作为时间函数的空间轨迹；（2）占据不同空间位置的物体的不同点共享相似的运动和外观模式；（3）物体的运动和外观模式常常可以相互预测，并且这两者在物体在时空中的形状和存在方面保持一致——从运动中估计出来的东西也可以从外观中估计出来。因此，外观和运动可以相互提供监督信号，并协同工作。基于不同类型的信息（如运动和外观），沿着不同的预测路径寻找一致的想法并进一步扩展，我们提出了一个新颖的视觉故事概念，旨在为无监督学习提供一种通用的方法。我们再次理解到，只有通过相互的协议（agreements），我们才能最终在野外实现无监督学习，在那里，由于空间和时间的结构、连贯性和一致性，这种协议会自然发生。

接下来，我们提出了新的空间和时间图结构，其中分割被描述为带有运动和外观约束的光谱聚类问题。分割被发现作为特征-运动矩阵的主要特征向量。从数学上讲，这种方法与我们早期的图匹配公式直接相关，该公式也利用了这样一个事实，即具有非负元素的对称矩阵的主特征向量自然地反映了其关联图的主要的、最强的聚类。从概念上讲，视频中的每个像素都是我们图形中的一个节点，但我们从不显式地计算矩阵，因为这是无法计算的。相反，我们提供了一种有效的方法，它可以找到矩阵的主特征向量，而不必实际构造矩阵。

原则5的应用：谱图匹配

原则5：偶然的匹配很少见。当它们发生时，通常表示模型和图像之间的正确匹配。匹配可以是几何的，也可以是基于外观的，虽然很少见，但当它们发生时，会形成一个强大的集群，以多种方式相互加强。

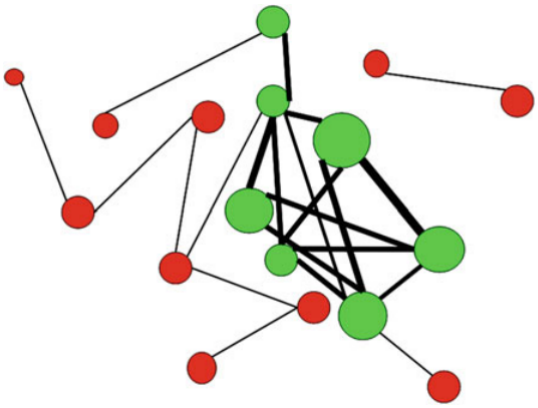
在这里，我们简要介绍了光谱图匹配算法，这一旨在给出图形匹配问题的近似解。该算法可以用于多种应用，因为它以最一般的形式处理图匹配问题，如整数二次规划，并且它不对一元和二阶项强加特定的数学形式。我们唯一需要的约束是一元和成对的分数必须是非负的，并且它们应该随着匹配的质量而增加，也就是说，随着候选匹配的变形误差的减小而减小。这些分数可以在外观和几何关系层面捕捉任何类型的变形/变化/错误。应用于计算机视觉的图匹配问题中，我们对图匹配问题的关键见解是，正确的匹配之间将有强有力的二阶分数，而不正确匹配之间不太可能有如此大的分数，因为偶然的几何匹配是非常罕见的事件。



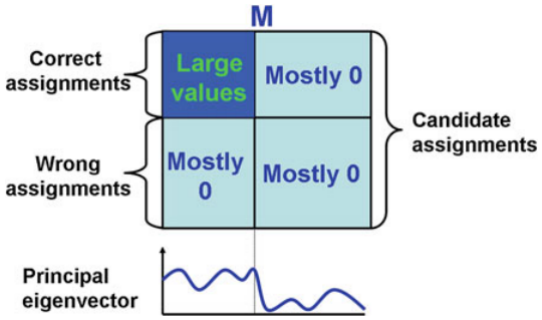
> 图1.6 成对的错误匹配（红色）不太可能保留几何图形，因此成对得分较低。正确的匹配将保留几何图 形，并且得分很高。二阶几何关系通常比基于局部外观的一元项更强大。在这个例子中，只考虑局部信息是不可能进行正确匹配的。只有正确的匹配才会保留成对的几何图形，因此鼓励使用这种类型的二阶几何信息来进行鲁棒匹配。

这些二阶分数的概率性质赋予了包含二阶项/分数的匹配矩阵M一个非常具体的结构：一个由正确的匹配组成的大值的强块，而其他地方几乎都是零值。这允许我们在优化步骤中放弃对解的整数约束，并只在之后施加整数约束，作为一个二值化过程应用于这个矩阵M的主特征向量。在图1.6中，我们提供了一个例子来说明算法背后的直觉。红色汽车的图像分辨率很低。因此，每个特征的局部外观不足以区分两个连续帧之间的可靠匹配。然而，成对的几何图形在帧之间被很好地保留，所以很有可能保存该几何图形的成对匹配是正确的。这只是一个演示示例，但使用成对几何约束进行鲁棒匹配适用于很多计算机视觉应用。

我们可以把M看作候选匹配图的加权邻接矩阵。每个候选匹配（或对应关系）(i, a)可以看作这个图中的一个节点，其包含了候选匹配之间局部外观一致性的信息。节点之间的链接可以包含关于候选匹配之间的成对几何信息保存得如何的信息。图1.7显示了这样一个候选匹配图的可能实例。较大的节点对应较强的一元分数（在局部外观水平的一致性较好），较粗的边对应与更大的成对分数，反映了在二阶几何水平上更强的一致性。我们期望正确的匹配将形成一个强连通簇，这个簇可以通过分析匹配图的加权邻接矩阵M的主导特征向量来找到。特征向量的元素可以解释为每个候选匹配都是正确的的置信度，利用整数约束对特征向量进行二值化，最终得到图匹配问题的近似解。在图1.8中，我们展示了矩阵M的可能结构。正确的匹配会在这个矩阵中形成一个含有大量成对元素的强块，而错误匹配之间的成对分值大多为零。这将反映在M的主特征向量中。



> 图1.7 正确的匹配（绿色显示）很可能通过建立更强的成对二阶协议（更粗的边）和更好地保留局部外观的特征（更大的节点）来形成一个强大的集群。



> 图1.8 矩阵M的结构：正确的匹配会在M中形成一个含有大量成对元素的强块，而错误匹配之间的成对分值大多为零。M的这个统计性质将反映在它的主特征向量上。

谱图匹配算法总结如下：

1. 创建候选匹配图：构建候选匹配ia的矩阵M，该矩阵将来自模型的特征i与来自图像的特征a对应起来，使非对角元素$M_{ia;jb}$获取候选匹配ia和jb之间的一致程度：它们有效地测量了来自模型的对(i,j)的几何形状和外观与来自图像的对(a,b)的几何形状和外观的匹配程度。该矩阵应该有非负元素，其值随着候选匹配ia和jb之间匹配（一致性）质量的增加而增加。因此，每个候选匹配ia都成为候选匹配图中的一个节点。$M_{ia;jb}$的边上的值衡量候选匹配ia和jb之间的一致性。
2. 高效优化：用幂迭代算法计算M的主导特征向量v。在实践中，少数迭代(10-20次)就足够了。特征向量的大小将与候选匹配的数量相同，并且只有非负值。每个v_{ia}的值表示候选匹配属于匹配图中主要、最强聚类的程度。置信值是对匹配ia正确的可能性的度量。
3. 找到最终的离散解：最终的离散解以贪婪的方式快速得到：(1)我们从还没有被选中或淘汰的ia中选择v_{ia}值最高的匹配ia^*；（2）然后，我们舍弃剩余的候选匹配中所有与ia^*冲突的匹配（可能具有相同的源或目标特征。然后我们回到第一步。

   这一过程一直持续到所有匹配图都被正确选择或丢弃。

原则7的应用：多代无监督分割学习

原则7：为了改进或学习新的类，我们需要增加训练数据的数量和难度，以及增加分类器的功率和多样性。通过这种方式，我们可以通过使用现有分类器之间的聚合来作为对新分类器的教师监督信号，以无监督的方式进行几代分类器的学习。

现在，在我们的导论章的结尾，我们结合了到目前为止所提出的思想，并介绍了一个系统，该系统在无监督的方式下学习，在多代师生中分割单个图像中的对象。它的核心结合了本章讨论的许多关键思想，特别是无监督学习的原则7。

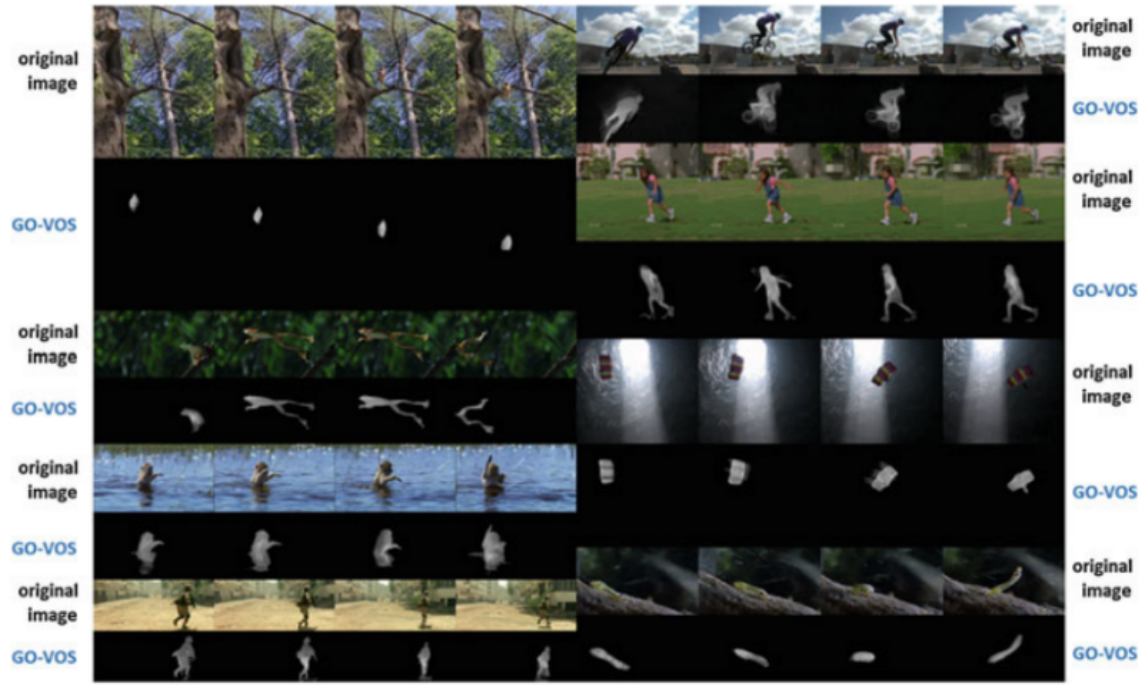
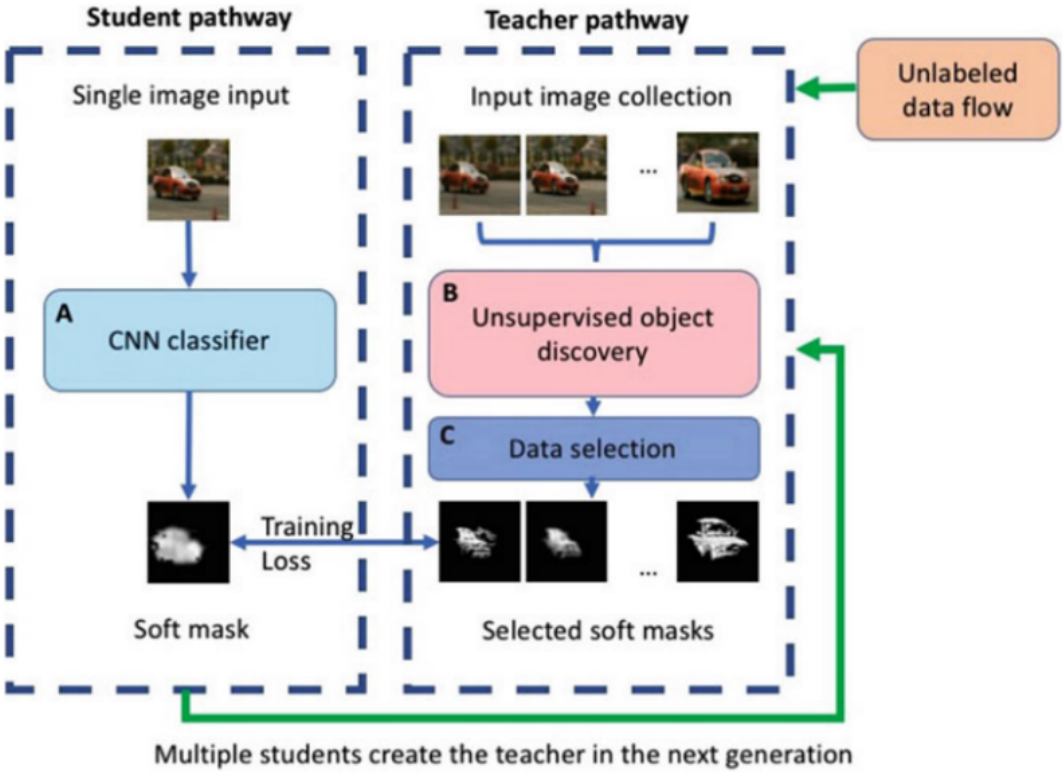


Fig. 1.14 Visual qualitative results on SegTrack dataset with our unsupervised GO-VOS formulation

> 图1.14 使用无监督GO-VOS公式在SegTrack数据集上获得可视化定性结果

这里介绍的系统由两条主要分支组成，一条是在视频或大型图像集合中执行无监督的对象发现——教师分支，另一条是学生分支，它从教师分支那里学习如何在单个图像中分割前景对象。这一无监督的学习过程将在几代学生和教师分支中继续进行。在算法1.6中，我们给出了此方法的高级描述。我们的方法确保了一代接一代的性能提高，其关键方面是：（1）存在一个无监督的选择模块，该模块能够拾取教师分支生成的高质量掩码，并将它们传递给下一代学生分支进行训练；（2）训练具有不同体系结构的多个学生，通过他们的多样性来帮助为下一次迭代训练培养更好的选择模块，在下一次迭代中与选择模块一起形成更强大的教师分支；（3）访问大量（可能是更复杂的）无标签数据，随着每一代变得更强大，这一步骤变得更加有用。

我们的方法是通用的，因为学生或教师的路径不依赖于特定的神经网络架构或实现。通过大量的实验和与最先进的方法的比较，我们还发现，这一系统适用于计算机视觉中的不同任务，如视频中的对象发现、无监督图像分割、显著性检测和迁移学习。



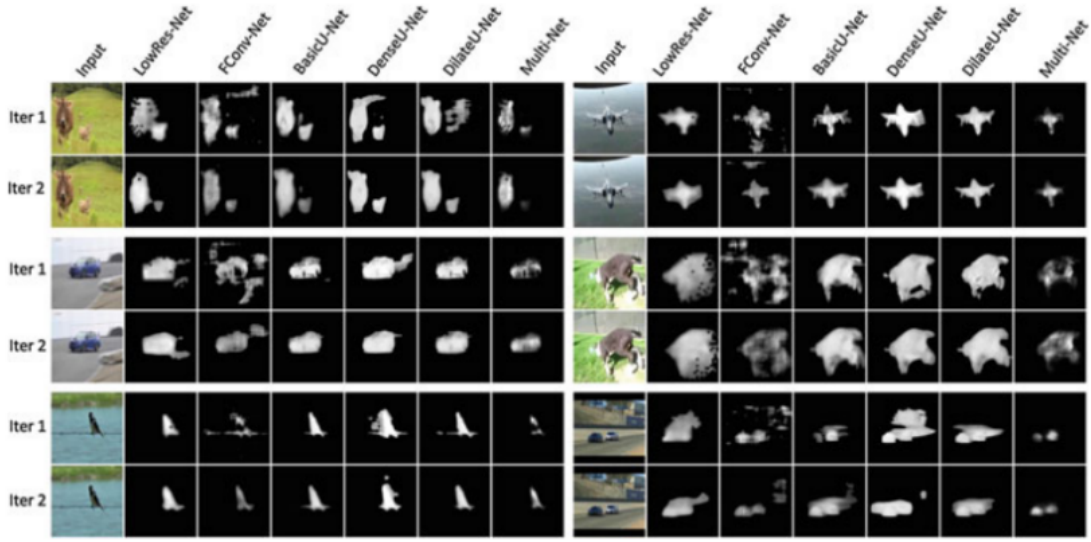
> 图1.15 双学生-教师系统被提出用于无监督学习以分割图像中的前景对象，其功能如算法1.6所示。它有两条路径：沿着教师分支，视频或大图像集合中的对象发现器（模块B）检测前景对象。然后根据无监督的数据选择过程（模块C），过滤得到软掩模。最终得到的成对图像——输入图像（或视频帧）和该特定帧的软掩模（充当无监督标签）用于训练学生路径（模块A）。整个过程可以重复多代。每一代训练几个学生CNN，然后由他们共同训练一个更强大的选择模块C（由深度神经网络建模)）并在整个算法的下一次迭代时形成一个整体更强大的教师路径。

在图1.15中，我们展示了整个系统的图形概述。在无监督训练阶段，学生网络（模块A）从无监督的教师路径（模块B和模块C）一帧一帧地学习，以在单一图像中产生相似的对象掩模。模块B发现图像或视频中的对象，模块C模块B生产的足够好的掩模，将其传递给模块A进行训练。因此，学生分支试图为模块C选择的帧模仿模块B的输出。学生分支的输入只是单个图像——当前帧——而教师分支可以访问整个视频序列。模块A的能力取决于模块B的性能。然而，正如我们在实验中看到的，选择模块C的能力促成了一个事实，即新学生分支的性能将超越其最初的教师模块B。因此，在整篇文章中，我们把B称为初始的"教师"，把B和C一起称为完整的"教师路径"。算法1.6中提出的方法遵循系统从一个迭代（代）到下一个迭代学习的主要步骤。

在算法1.6的第一次迭代中，无监督教师（模块B）可以随时访问信息——一段视频。相比之下，学生在结构上更深入，但它只能访问一个图像——当前的视频帧。因此，教师及时发现的信息被学生在抽象的神经层上更深入地捕获。在第一次迭代中训练了几个具有不同架构的学生网络。为了仅使用高质量掩模作为监控信号，应用了无监督掩模选择程序（模块C）（在迭代1中非常简单）。一旦训练了几个学生网络，它们就可以在下一次迭代中（以各种方式）形成教师路径，以及一个更强的无监督选择模块C，如深度神经网络EvalSeg-Net。简单来说，EvalSeg-Net学会预测不同学生之间的输出掩模的一致性，这在统计上发生在掩模质量较好的时候。因此，EvalSeg-Net可以作为一种无监督的掩模评估程序和强选择模块。然后，在下一代，我们在更大的一组无标签的视频或图像集合上运行新形成的教师路径（模块B和模块C），为下一代学生产生监督信号。在实验中，我们展示了模块B和C在下一个迭代时的改进，以及数据量的增加，这对于在下一代提高精度来说都是很重要的，尽管不是全部必要的。

请注意，虽然在第一次迭代中，教师路径需要接收视频序列作为输入，但从第二代开始，它也可以接收大型图像集合作为输入。由于训练期间需要非常高的计算和存储成本，我们将实验限制在两代的学习时间内，但我们的算法是通用的，可以运行多次迭代。大量的实验表明，在视频和图像中，即使两代也足以在对象发现方面超越当前的技术水平。我们还以实验的方式证明，从一代到下一代，每一个相关的组成部分，包括学生个体（模块A）、教师（模块B）和选择模块C，都有切实的改善。

在图1.16中，我们展示了在不同迭代中用作学生网的不同模型之间的视觉比较结果，包括不同模型的集合。从一次迭代到下一次迭代的质量改进是显而易见的，而且平均精度（F-measure）提高了5%也反映了这一点。下面，我们在此重复一些我们在分析结束时所作结论和观察。



> 图1.16 在每次迭代时模型之间的可视比较。Multi-Net，如图所示，表示五种模型之间的像素乘法。请注意，第二代学生的掩模效果更好，形状更好，孔更少，边缘更锋利。还要注意的是，Multi-Net集合的召回率相对较差，会产生更小的、被侵蚀的掩模。

算法1.6 前景对象分割的无监督学习

步骤1：沿教师路径在未标记的视频（或之后迭代得到的图像集合）上执行无监督的对象发现迭代

步骤2：自动过滤掉在上一步生产的劣质软掩模

步骤3：使用剩下的掩模作为监督信号，沿学生路径训练一个或多个学生网

步骤4：用当前一代的一个或几个学生网（一个新的模块B）作为新教师，并学习一个更强大的软掩模选择器（一个新的模块C），为下一次迭代作准备

步骤5：扩展未标记的视频或图像数据集，并返回到步骤1来训练下一代（注意：从第一次迭代开始，训练数据集也可以用未标记图像的集合来扩展，而不仅仅是视频）