# 第1章

# 绪　论

## 1.1　视觉心智计算的概念

什么是心智(mind)？英文“mind”是一个多义词，相对于物质或身躯时可以指心脏；在谈论理智时可以指健全的心智、正常的神志；也可以用作智力、理解力、想法等概念；在哲学领域“mind”一般译作心灵；在智能科学领域“mind”被译作心智，是指一系列认知能力组成的总体，包括情感、意志、感觉、知觉、表象、学习、记忆、思维、直觉等。这些能力赋予个体意识，使其拥有思考的能力，可以做出判断并记忆事物。

心智理论(theory of mind，ToM)是一个心理学术语，表示一种能理解自己及他人心理状态的能力，这里的心理状态包括情绪、信仰、意图、欲望等。心智理论是心理学、认知科学和神经科学领域的重要研究对象，被广泛认为是社会认知的关键组成部分，在社交互动中扮演着至关重要的角色，使人们能够预测和解释他人的行为，有效地进行交流及合作。

康康爱吃草莓(图 1-1)等系列实验是 ToM 测试中一个直观的实验范式，通过一系列实验，证明了典型发育儿童和孤独症儿童在心智理论发展方面存在差异。

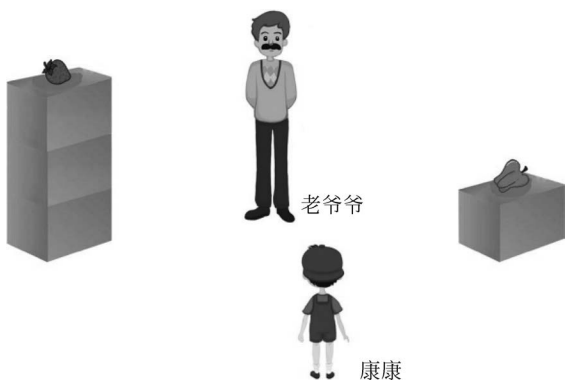


图 1-1　康康爱吃草莓[1]

实验的关键点在于屏幕中间：一个实验条件中间是老爷爷，另一个实验条件中间是一棵树。而康康想吃的草莓在高处的箱子上。如果康康想吃到草莓，必须得到老爷爷的帮助才能实现，这就涉及社会认知的问题。而树不是生命体，无法为康康提供帮助。儿童需要站在康康的角度，根据实验中的提示信息，预测康康的意图，这一过程涉及视角的转换，是心智理论的重要部分。实验结果也表明，典型发展儿童会更多地看向老爷爷，即他们知道康康需要向老爷爷寻求帮助，而孤独症儿童则没有表现出相同的反应模式，表明典型发展儿童可以进行心智理论的推理，可以预测他人的心理状态和行为。

进一步地，对心智理论进行数学建模以实现可计算的过程，称为心智计算。心智计算是对心理符号的计算，是模拟大脑进行信息加工的过程。建立心智模型的目的是探索和研究人的思维机制，特别是信息处理机制，同时也为设计相应的人工智能系统提供新的体系结构和技术方法。举例来说，一个著名的心智理论计算模型是贝叶斯心智理论（Bayesian theory of mind，BToM），由 Baker、Saxe 和 Tenenbaum 在 2009 年提出[2]。

BToM 是一个 ToM 的计算框架，即人类推理智能体的心理状态（如信念和期望）的能力。BToM 将心智理论的核心，即对信念和欲望相关行动的预测模型，表述为部分可观测马尔可夫决策过程，并根据对某个环境背景下智能体行为的观测结果，利用贝叶斯推理重建智能体的联合信念状态和奖励函数。实验向参与者展示了简单空间场景中智能体的移动序列，并要求参与者对智能体的期望及环境中其他未观察到的其他部分进行联合推断，以此来测试 BToM。其中，观察者对世界的表征由环境状态和智能体状态组成（图 1-2）。BToM 已被用于解释广泛的社会认知现象，包括错误信念推理[3]、观点采择[4]和社会学习[5]。它也被用于研究精神疾病。

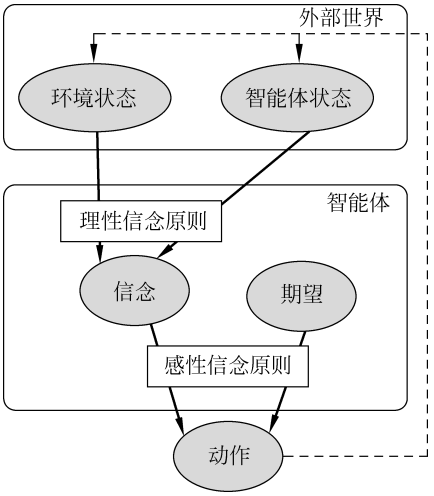


**图 1-2　心智理论的示意模型**

视觉心智是心智理论概念的一个具体方面，指的是通过视觉认知过程，如面部表情、身体语言和其他视觉线索，理解他人心理状态的能力。理解他人心理状态的能力很大程度上取决于解读非语言线索的能力，例如推断他人的情绪、意图和信仰等。研究表明[3]，视觉心智能力在生命早期就开始发展，12 个月大的婴儿就具有一定的推理能力，能够从目的论的角度解释行为，即将行为视为达成目标的手段。随着年龄的增长，解读非语言线索的能力变得更加复杂，使人们能够对他人的心理状态做出越来越细致的推断。

机器可以思考吗？或者说心智本身会是一台思考的机器吗？实现机器思考的关键是人们是否可以实现心智计算。近年来计算机技术的迅速发展改变了人们对这些问题的观点，这些核心技术使人们有望研发出具备感知、语言理解、模仿推理、决策以及其他心理过程的智能机器。

本书将视觉心智计算描述为使用数学和计算机技术等工具，模拟和解释涉及理解他人心理状态的视觉认知过程。常用的工具方法包括贝叶斯推理、强化学习等。贝叶斯模型可以通过先前的行为或社会背景信息推测出一个人可能的目标或意图。当新的信息，例如该人的行动或言语提示，不断出现时，模型会更新其信念，并推断出更多关于该人心理状态的信息。视觉心智计算还可以利用神经网络或其他机器学习技术来模拟社会推理中涉及的认知过程。例如，一个模型可以对大量社交互动的数据进行训练，并使用强化学习算法，学习如何准确预测他人的心理状态。举例来说，在一场社交聚会上，观察到一个人多次看向门口。在这种情况下，根据贝叶斯模型可以由该人的行为推断出他可能在等待某人。进一步观察到他不断查看手机，模型的信念会更新，进一步支持他在等待某人的推断。同样地，视觉心智计算模型可以通过分析大量类似的社交互动数据，学习到在上述情境中，频繁看向门口并查看手机通常代表着等待某人。通过这种方式，模型能够模拟和预测他人的心理状态。

总的来说，这些计算模型是研究心智理论的有力工具，因为它们允许研究人员测试和改进不同的社会认知理论。然而，这些模型仍然是对理解他人所涉及复杂认知过程的简化抽象，可能无法完全捕捉人类社交互动的丰富性和多样性。视觉心智理论（V-ToM）由 Peng Zhou、Huimin Ma 和 Bochao Zou 等在 2023 年提出[1]。V-ToM 成功地将 ToM 操作成细粒度框架，提出基于心智理论的细粒度模型。该模型以心智理论为框架，研究心智理论的两个核心组成部分：情感性和认知性心智理论。通过构建具有情感和认知意义的视觉场景刺激，明确描述人类对他人的信仰和情感进行推断的四阶段过程，即视觉处理、心理加工、评估和转移实施。本书第 3 章将对 V-ToM 模型进行详细描述。

## 1.2 视觉心智计算的内容

视觉心智计算的主要内容是人们利用视觉线索信息推断他人的心理状态，例如他们的信念、欲望、意图和情感等。因为视觉线索通常能够提供有关他人意图和

信念的信息，而视觉心智计算就是对人们通过视觉信息推断他人心理状态的过程进行建模[2,6]。

视觉心智计算的内容根据研究者采取的具体方法而异。然而，一般而言，视觉心智计算涉及以下内容的研究。

(1) 视觉感知。模型需要包含视觉感知算法，以便检测和分析视觉线索，例如注视方向、面部表情和动作姿态。

(2) 心理状态的表征。模型需要以一种可操作的方式表示心理状态，例如信念、意图和欲望，并能从视觉线索中进行推断。

(3) 推理机制。模型需要使用推理算法，基于视觉线索和其他可用信息，来预测他人的心理状态。

(4) 学习和适应。模型可能包含学习机制，以适应新的环境和情况，并从反馈中学习。

(5) 与其他认知过程的集成。模型可能需要与其他认知过程(例如语言处理、记忆和注意力)相结合，以支持更复杂的社会认知。

总的来说，视觉心智计算的内容高度依赖所研究的具体问题和研究者采用的方法。一个成功的模型需要整合多个组成部分，以准确捕捉社会认知和行为的复杂性。视觉心智计算的过程包括以下步骤：首先，观察他人的行为和动作，利用这些信息形成关于他们心理状态的假设；其次，通过捕捉到的视觉线索，如他们的凝视方向、面部表情、身体姿势以及其他非语言信号，进一步细化这些假设，从而更准确地推断他们的心理状态；最后，利用情境信息，如对情境或社交规范的了解，进一步细化对他人心理状态的推断。

视觉心智理论的计算模型利用数学和统计方法，模拟人们使用视觉线索和感知信息推断他人的心理状态。这些模型有助于解释人们如何理解他人的心理状态，并在社交场合中做出预测。视觉计算心智理论是利用视觉线索推断他人心理状态的过程，对社会认知和交互起着至关重要的作用。

## 1.3 视觉心智计算的发展

心智理论研究在过去的几十年中，逐步从解决特定心智问题的尝试，演变为不同风格的第一代理论和第二代理论：以表征-计算为核心的心智计算理论(the computational theory of mind，CToM)和以具身性观念为理论特征的第二代心智理论。

20 世纪 70 年代，纽厄尔(Newell)和西蒙(Simon)提出了物理符号系统理论，主张认知过程可以通过一组符号及其相关规则的物理操作来模拟和实现；随后，经由普特南(Putnam)、马尔(Marr)、福多(Fodor)、派利夏恩(Pylyshyn)等学者的发展，通过强调认知系统的功能性特征、信息处理和符号操作，奠定了第一代心智

理论的基本内核，即以表征-计算为核心的心智计算理论。这一理论框架中最关键的预设是：认知过程是对人们周围世界心理表征的生成、转换和删除的心理操作，认知状态则是内部心理表征之间的关系，这些关系、表征和操作都可被视为计算的过程，人脑心智系统可被当作一台“计算机”。

1975年，福多提出思维语言假设后，心智计算理论演变为包含符号计算(digital computational theory of mind，DCToM)和连接计算(connectionist computational theory of mind，CCToM)两种形式。后续在德雷福斯(Dreyfus)和塞尔(Searle)等对强人工智能激烈批判的刺激下，以具身性观念为理论特征的第二代心智理论逐渐登上历史舞台。其中，具身性观念认为人类学习并理解世界是通过人体感官和身体在空间中的移动和互动实现的，而不仅仅是被动地接收信息。认知科学家加拉格尔(Shaun Gallagher)以具身认知(embodied cognition)、嵌入认知(embedded cognition)、延展认知(extended cognition)和生成认知(enactive cognition)概括了这一理论的核心理念，有时称为4EC。其中，具身认知强调身体在认知过程中的重要性，它认为人们的认知不仅发生在大脑中，而且与人们的身体密不可分。嵌入认知强调环境在认知中的作用。该理论认为，认知过程不仅限于个体的大脑和身体，而是深深地嵌入与环境的互动中。延展认知的概念进一步拓展了认知的范围，提出人们的认知系统可以超越个体的身体，延展到外部的工具和技术中。生成认知强调认知是一个动态生成的过程，通过行动和交互与世界共同创造。

心智表征作为心智理论的核心概念之一，涵盖信息或知识在心理活动中的表示和记录方式。这种表征可以看作外部事物在心理活动中的内部再现，一方面它反映和代表客观事物，另一方面又是心理活动进一步加工的对象。信息或知识以符号的形式存在，比如文字和数字，这些符号被赋予了特定的意义，因此具有一定的价值。心智表征不仅是对知识的形式化描述，也是一系列关于知识描述的约定，构成一种心理活动可接受的数据结构。常见的心智表征方式包括逻辑、产生式系统(规则)、框架等。

以图式为例，它实质上是一种心理结构，是能帮助人们知觉、组织、获得和利用信息的认知结构。认知心理学家认为，人们在认知过程中通过对同一类客体或活动基本结构的信息进行抽象概括，在大脑中形成的框图便是图式。例如，一个孩子第一次形成关于马的图式时，他知道马是大的，有毛发、四条腿和尾巴。但当他遇到奶牛时，可能会错误地称其为马。在被告知这是一种不同的动物后，孩子会修改对马的图式，并为奶牛创建一个新的图式。通过这样的经历，孩子逐渐意识到，虽然一些马是非常大的动物，但另一些马可能很小。皮亚杰(Piaget)、鲁梅尔哈特(Rumelhart)等认为，图式由表示概念要素的若干变量组成，是一种知识框架及分类系统。1987年，纽厄尔和莱德、罗森勃卢姆提出了一个通用解题结构SOAR(state，operator and result)，希望把各种“弱方法”(基于启发式的解题方法)都实现在这个解题结构中。SOAR是一种理论认知模型，它既从心理学角度对人类认知

建模，又从知识工程角度提出一个通用解题结构。它模拟了人类的解题过程，通过状态、操作符和结果来描述和执行解题步骤。SOAR 旨在整合多种弱方法，使系统能够在不同问题情境下选择和应用适当的操作符，逐步解决复杂问题。

如上所述，视觉计算的心智理论是指使用视觉线索和感知信息来推断他人的心理状态。这种心智理论的方法基于人们使用视觉信息推断他人的心理状态，如信念、欲望和意图。但解释或理解、预测他人行为的心理基础机制是什么？在理论层面有如下不同的解答。

(1)“社会脑假说”(the social brain hypothesis，SBH)，由 Dunbar 于 1998 年提出[7]。SBH 假设灵长类动物大脑的演化是由社交认知的需求驱动的，包括理解他人的心智状态。基于 SBH 的计算模型已被用于模拟灵长类动物与其他动物社交认知的演化[8]。

(2)“理论之理论”(theory-theory，TT)，由戈普尼克(Gopnik)和梅尔佐夫(Meltzoff)于 1997 年提出[9]。TT 假设儿童通过自己的经验和观察他人来发展心智理论。基于 TT 的计算模型已被用于解释儿童如何获得信念、欲望和情感等[10]。

(3)“模拟理论”(simulation theory，ST)，由戈登(Gordon)于 1986 年提出[11]。ST 假设人们通过在自己的心智中模拟心智状态理解他人的心智状态。基于 ST 的计算模型已被用于解释人们如何推理情感、意图和错误的信念[12]。

以上每个模型都提出了人们如何推理他人心智状态的独特观点，并被广泛用于解释各种社交认知现象。

近年来，在理论假说发展的基础上，不断有学者采用多种实验手段或工具进行心智理论的研究。如其中一项研究调查了儿童如何利用面部表情推断他人的情绪，该研究发现，儿童由面部表情推断情绪的能力随着年龄的增长而提高，并且这种能力与他们的心智理论能力有关[13]。另一项研究通过眼动跟踪技术探究了儿童和成人如何利用注视方向推断他人的心理状态。结果显示，成年人和儿童都利用注视方向推断意图，但儿童的准确度低于成年人[14-15]。还有研究调查了视觉工作记忆和心智理论之间的关系，发现在视觉工作记忆任务中表现更好的参与者更擅长推断他人的心理状态[16]。这些研究提供了利用视觉线索和感知信息推断他人心理状态的方法，以及这些推断能力在童年和成年阶段的发展情况。

在计算模型方面，有研究将视觉信息与先前的知识相结合，以推断他人的心理状态，并在虚拟现实场景中进行测试[17]，其参与者需要推断一个角色的意图。结果表明，基于视觉线索，该模型在预测角色的意图方面表现出色。如果将视觉信息与语言处理相结合构建计算模型，并使用模型推断故事中角色的心理状态，计算模型能够高精度地预测参与者的回答[18]。还有研究构建神经网络理论推理模型，使用视觉特征预测他人的心理状态。通过注视方向和面部表情预测他人的心理状态[19-20]。

这些研究表明，计算模型有潜力将视觉信息与其他知识源相结合，以推断他人的心理状态。未来的研究可能会开发更复杂的模型，以更好地捕捉理论推理过程的复杂性。这种综合方法可能会为人们提供更深入的理解，更准确地预测和解释人们在社交互动中的行为和心理状态。

## 1.4　视觉心智计算的应用

视觉心智理论计算模型在许多领域都有潜在的应用，如人机交互、教育、诊疗精神心理疾病等。在人机交互领域，视觉心智理论的计算模型可以用于改善人类和机器人之间的社交互动，如具有心智理论计算模型的机器人可以更好地理解和回应人类的意图和情感，从而实现更自然、更吸引人的交互。心智理论计算模型也可以在教育环境中使用，通过了解如何使用视觉线索推断他人的心理状态，学生可以更好地理解同伴的观点，提高他们的社交互动，帮助学生发展社交技能和情商。在心理学和精神病学方面，心智理论模型可以用于更好地理解和治疗各种心理和精神障碍，如孤独症谱系障碍、精神分裂症和边缘型人格障碍，通过了解这些障碍的患者如何处理与心智理论相关的视觉信息，研究人员和临床医生可以开发更有效的筛查及治疗方法。在市场营销和广告方面，心智理论计算模型可以用于市场营销和广告，以更好地了解消费者的观点和动机，通过分析消费者的视觉线索和其他行为，营销人员可以更好地定制他们的产品信息，以吸引目标受众。

总的来说，视觉心智理论的计算模型具有广泛的应用，并且有潜力增进人们对社会认知和行为的理解。

## 1.5　视觉心智计算与人工智能

视觉心智计算与人工智能之间存在紧密关系，因为心智计算旨在开发能够理解人类社交认知和行为的算法和模型。具体而言，视觉心智计算致力于使机器从他人的面部表情、肢体语言和凝视方向等视觉线索中推断他人的心理状态，是人工智能研究的重要领域之一，因为它旨在使机器能够以更自然、更类似于人类的方式理解人类并与之交互。视觉心智计算在人工智能领域有着广泛的潜在应用，其中一个重要方面是发展具有社交智能的机器人。为机器人配备心智理论模型，可使其更好地理解和响应人类的意图和情感，从而实现更自然、更吸引人的交互。总的来说，视觉心智计算的发展代表着人工智能研究的一个重要领域，具有广泛的应用前景。

具身智能是人工智能的重要分支。具身智能系统不仅包括脑部的符号处理，还包括机器或生物体身体与环境之间的交互作用。其强调系统通过感知、行动和互动实现智能表现，而不仅仅是简单地处理符号或信息。这种系统利用感知机制

获取信息，通过身体执行动作影响环境，并根据环境反馈做出适应性调整。具身智能系统的核心思想是将智能看作是与身体、环境和经验相结合的过程，而不仅仅是抽象的符号处理。

视觉心智计算可以帮助机器理解人类的情感和意图，而具身智能系统则可以使机器人更自然地与环境和人类交互，使其表现出更灵活、更智能的行为，共同推动人工智能技术向着更智能、更自然和更人性化的方向发展。

## 1.6 本书的内容结构

本书的目的是系统地介绍视觉心智计算这一相对新兴研究领域的基本内容，包括视觉基础、视觉心智理论、视觉计算理论等理论基础章节，心智计算建模等前沿方法章节，以及视觉心智计算应用实例章节。

其中，第 2 章介绍视觉认知的概念、神经基础、认知基础理论及视觉与注意、记忆、学习、决策等核心认知能力的关联；第 3 章阐述心智理论的概念、相关研究进展，探讨心智理论与人工智能的相关性，并提出视觉心智理论模型；第 4 章阐述视觉计算方法，以马尔计算视觉为基础，讲解视觉计算的经典方法与最新进展，介绍视觉心智计算中的视觉计算推理与决策相关内容；第 5 章是视觉心智计算的核心方法部分，介绍心智计算建模方法，包括心智理论建模、智能体心智建模，并讨论深度学习和心智理论的差距；第 6 章为应用章节，面向心理状态评估与自动驾驶具体任务介绍视觉心智计算理论方法的实际应用。

## 参考文献

[1] ZHOU P, ZHAN L, MA H. Predictive language processing in preschool children with autism spectrum disorder: An eye-tracking study[J]. Journal of Psycholinguistic Research, 2019, 48: 431-452.

[2] BAKER C L, SAXE R, TENENBAUM J B. Action understanding as inverse planning[J]. Cognition, 2009, 113(3): 329-349.

[3] LIU S, ULLMAN T D, TENENBAUM J B, et al. Ten-month-old infants infer the value of goals from the costs of actions[J]. Science, 2017, 358(6366): 1038-1041.

[4] AICHHORN M, PERNER J, KRONBICHLER M, et al. Do visual perspective tasks need theory of mind?[J]. Neuroimage, 2006, 30(3): 1059-1068.

[5] LUCAS C G, GRIFFITHS T L, XU F, et al. The child as econometrician: A rational model of preference understanding in children[J]. PLOS ONE, 2014, 9(3): e92160.

[6] MARGOLIS E, SAMUELS R, STICH S P. The Oxford handbook of philosophy of cognitive science[M]. Oxford: Oxford University Press, 2011.

[7] DUNBAR R I M. The social brain hypothesis[J]. Evolutionary Anthropology Issues News & Reviews, 1998, 6(5): 178-190.

[8] DUNBAR R I M. The social brain hypothesis and its implications for social evolution[J]. Annals of Human Biology,2009,36(5): 562-572.

[9] GOPNIK A, MELTZOFF A N. Words, thoughts, and theories[M]. Cambridge: Mit Press,1998.

[10] GEISLER W S,DIEHL R L. A Bayesian approach to the evolution of perceptual and cognitive systems[J]. Cognitive Science,2003,27(3): 379-402.

[11] GORDON R M. Folk psychology as simulation[J]. Mind & Language, 2010, 1(2): 158-171.

[12] HARRIS P L. From simulation to folk psychology: the case for development[J]. Mind & Language,1992,7(1/2): 120-144.

[13] FITZPATRICK P, FRAZIER J A, COCHRAN D, et al. Relationship between theory of mind, emotion recognition, and social synchrony in adolescents with and without autism [J]. Frontiers in Psychology,2018,9: 1337.

[14] EINAV S,HOOD B M. Children's use of the temporal dimension of gaze for inferring preference[J]. Developmental Psychology,2006,42(1): 142.

[15] KANNGIESSER P,ITAKURA S,ZHOU Y,et al. The role of social eye-gaze in children's and adults' ownership attributions to robotic agents in three cultures[J]. Interaction Studies,2015,16(1): 1-28.

[16] MUTTER B, ALCORN M B, WELSH M. Theory of mind and executive function: Working-memory capacity and inhibitory control as predictors of false-belief task performance[J]. Perceptual & Motor Skills,2006,102(3): 819-835.

[17] CHEN X L,HOU W J. Gaze-based interaction intention recognition in virtual reality[J]. Electronics,2022,11(10): 1647.

[18] NARANG S,BEST A,MANOCHA D. Inferring user intent using Bayesian theory of mind in shared avatar-agent virtual environments[J]. IEEE Transactions on Visualization and Computer Graphics,2019,25(5): 2113-2122.

[19] GRAHAM R,LABAR K S. Neurocognitive mechanisms of gaze-expression interactions in face processing and social attention[J]. Neuropsychologia,2012,50(5): 553-566.

[20] BYOM L J,MUTLU B. Theory of mind: Mechanisms, methods, and new directions[J]. Frontiers in Human Neuroscience,2013,7: 413.

# 第2章

# 视觉基础

## 2.1 视觉认知的概念

认知(cognition)是指获取、处理、存储和使用信息涉及的心理过程。它包括广泛的心理活动,例如感知、注意、记忆、语言、推理等,如图 2-1 所示。认知涉及人类机能的方方面面,从基本的生存技能到复杂的智力活动,如科学发现和艺术创作等。为了更好地理解人类思维的运作方式、优化人类的心理过程,认知心理学、神经科学、语言学等领域都对人类认知进行了不同层面的探索[1]。

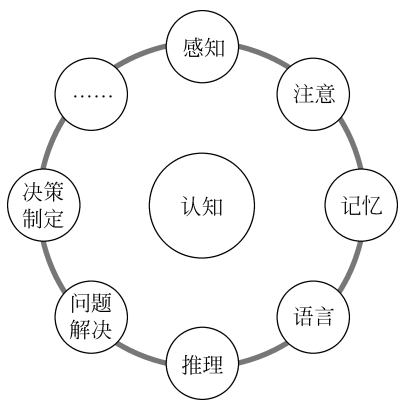


**图 2-1 认知包括的心理活动与过程**

视觉作为人类最重要的感官之一,在人类认知世界的过程中扮演着重要的角色。视觉认知(visual cognition)是研究人们如何感知、处理和解释视觉信息的学科。研究内容包括大脑如何处理视觉刺激、视觉注意力如何运作、视觉记忆如何运作,以及如何使用视觉信息做出决策并解决问题。视觉认知是一个多学科交叉领域,它利用心理学、神经科学、计算机科学和其他相关领域的多学科知识来探究视觉感知和认知背后的复杂过程[2]。视觉认知涉及注意、记忆、学习、决策等多个人类核心认知过程,本节将进行介绍。