图学学报
Journal of Graphics
ISSN 2095-302X,CN 10-1034/T

《图学学报》网络首发论文

题目：视觉图灵：从人机对抗看计算机视觉下一步发展
作者：黄凯奇，赵鑫，李乔哲，胡世宇
收稿日期：2021-04-06
网络首发日期：2021-05-08
引用格式：黄凯奇，赵鑫，李乔哲，胡世宇．视觉图灵：从人机对抗看计算机视觉下一步发展．图学学报.
https://kns.cnki.net/kcms/detail/10.1034.T.20210507.1635.004.html

网络首发时间：2021-05-08 14:47:00
网络首发地址：https://kns.cnki.net/kcms/detail/10.1034.T.20210507.1635.004.html

图 学 学 报
JOURNAL OF GRAPHICS

视觉图灵：从人机对抗看计算机视觉下一步发展

黄凯奇[1,2]， 赵 鑫[1]， 李乔哲[1]， 胡世宇[1]

(1. 中国科学院自动化研究所智能系统与工程研究中心，北京 100190；
2. 中国科学院脑科学与智能技术卓越创新中心，上海 200031)

摘 要：计算机视觉一直是人工智能研究的热点方向，经过近 60 年的发展，已经在算法、技术和应用等方面取得了巨大的进步。近十年来，以大数据、大算力为基础的深度学习进一步推动计算机视觉走向大模型时代，但其算法适应能力仍然和人类存在较大差距。本文从视觉任务评估评测（评测数据集、评测指标、评估方式）出发，对计算机视觉的发展进行了总结，对现存的依赖大数据学习的计算机视觉发展问题进行了梳理和分析，从人机对抗智能评测提出了计算机视觉下一步发展方向：视觉图灵。最后对视觉图灵发展方向进行了思考和讨论，探讨了未来研究可能的方向。

关 键 词：计算机视觉；视觉图灵；评估评测；图灵测试；数据集
中图分类号：TP 391
文献标识码：A **文 章 编 号**：2095-302X(2021)03-0000-00

Visual Turing: the next development of computer vision in the view of human-computer gaming

HUANG Kai-qi[1,2], ZHAO Xin[1], LI Qiao-zhe[1], HU Shi-yu[1]

(1. Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;
2. CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China)

Abstract: Computer vision has gained wide attention in the research of artificial intelligence. After nearly 60 years of its development, great achievement has been made in aspect of algorithms, technologies, and applications. Over the past decade, deep learning, which is on the basis of big data and huge computation power, has further ushered computer vision in an era of large model. However, there remains a huge gap between algorithm adaptability and human beings. From the perspective of visual task evaluation (in terms of datasets, metrics, and methods), this paper summarized the development history of computer vision. In addition, a systematic analysis was conducted on the existing problems and obstacles for the development of computer vision heavily dependent on big data learning. Based on the analysis, this paper argued that the visual Turing test could be the next research direction of computer vision. Finally, the development of the visual Turing test and its potential research were discussed.

Keywords: computer vision; visual Turing; evaluation of visual tasks; Turing test; datasets

1 绪论

计算机视觉旨在通过对人类视觉系统进行建模，让机器具备感知视觉信息的能力。作为人工智能技术的研究热点，计算机视觉技术经过近60年的发展，已经在理论方法、关键技术和实际应

收稿日期：2021-04-06；定稿日期：2021-04-21
Received：6 April, 2021；**Finalized**：21 April, 2021
第一作者：黄凯奇(1977–)，男，研究员，博士。主要研究方向为计算机视觉、模式识别、智能视觉监控、人的认知信息处理。E-mail：kqhuang@nlpr.ia.ac.cn
First author：HUANG Kai-qi (1977–), male, researcher, Ph.D. His main research interests cover computer vision, pattern recognition, intelligent vision monitoring, human cognitive information processing. E-mail：kqhuang@nlpr.ia.ac.cn

用等方面取得巨大进步[1-2]，并广泛应用于智慧城市、自动驾驶、智能医疗等领域。作为引领计算机视觉发展的风向标和催化剂，评估评测所采用的数据集、评测指标、评估方式的演变给整个计算机视觉研究的发展带来了多次大的变革。其中，随着大规模图像数据集ImageNet[3]发布，以大数据、大算力为基础的深度学习方法在人脸识别、物体检测、图像分割、目标跟踪等领域大幅度超越了传统方法的性能，引领计算机视觉发展到了依赖大规模计算方法的时代。

以无人驾驶为例，深度模型需要通过对周围环境的感知，完成对车辆运动的决策。以特斯拉为代表的科技公司已将具备自主泊车、自主变道、主动避障等功能的车辆进行量产，并完成在城市街道上的自动驾驶（autosteer on city streets）系统测试。该系统以 30 亿英里驾驶数据为基础完成算法的搭建[4]，然而当面对恶劣天气、复杂车流、障碍物干扰时，依赖于视觉传感器的自动驾驶系统仍然无法实现精准的感知和决策。2020 年 6 月，特斯拉 Model 3 因未正确识别横向侧翻的白色大货车，在高速公路上以 110 公里的时速与货车发生碰撞。这与人类在复杂场景甚至在对抗环境下的感知能力存在巨大的鸿沟。这类问题让人们对当前依赖大数据、大算力的计算机视觉发展模式产生思考和质疑，是什么原因导致这些方法在实验室环境下性能优异，但对真实应用场景的适应能力仍和人类的能力存在较大差距？计算机视觉发展可能的方向在哪里？针对以上问题，多位学者和专家从计算机视觉理论、方法、研究内容等开展了探讨，提出了许多有建设性的观点[5-7]。与此不同，本文从计算机视觉算法和技术应用出发，探讨以计算机视觉算法评估评测（评测数据集、评测指标、评估方式）为主要视角，对计算机视觉的发展历程进行梳理，并对各个阶段存在的问题进行分析，从而提出计算机视觉发展的下一步思考和建议。

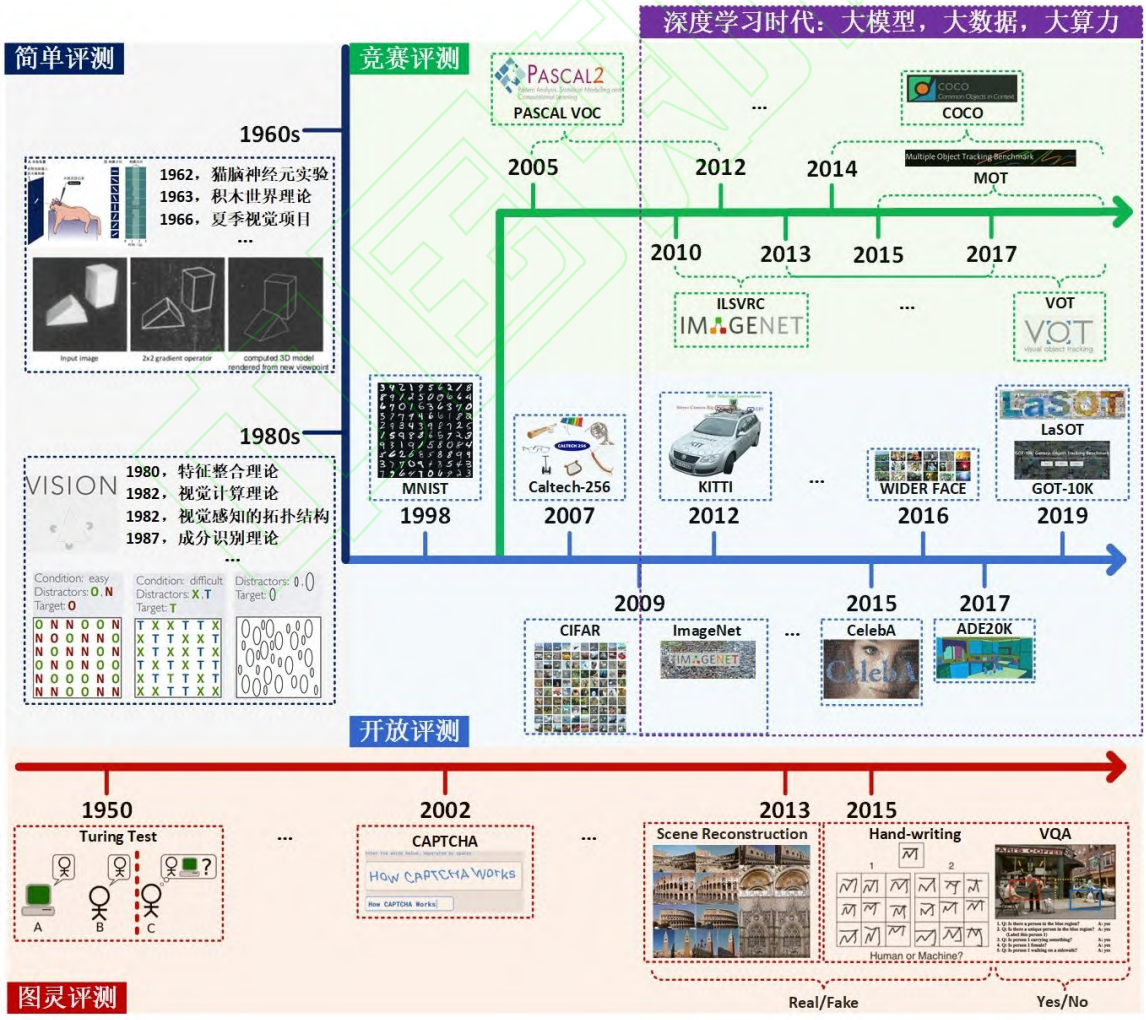

图 1　从视觉任务评估评测看计算机视觉发展

Fig. 1　The development of computer vision from the perspective of visual task evaluation

算法验证是计算机视觉算法实验的重要组成，算法的评估评测是计算机视觉理论之外的另一个重要部分。本文按照算法评估评测将计算机视觉发展划分为简单评测、开放评测、竞赛评测和图灵评测 4 个阶段（图 1）。早期，计算机视觉理论处于逐步完善阶段，相关实验在简单环境下依托少量数据完成对理论的验证。随着视觉理论和框架的逐步完善，其研究重点逐步细化到相关具体任务的研究，如物体检测、字符识别、人脸识别等，产生了包括数字手写识别数据集 MNIST[8]、图像分类数据集 CIFAR-10[9]与 CIFAR-100[10]等在内的系列数据集。为了更加公开公平地评测算法性能，不仅开放数据集，对评测指标也逐步统一，诞生了依托于竞赛的评测方式，如针对目标分类、检测和分割的 PASCAL VOC[11]竞赛、ImageNet[3]（ImageNet large Scale Visual Recognition Challenge，ILSVRC）大规模视觉识别挑战赛等，对于推动计算机视觉发展取得了巨大效果。然而，简单评测、开放评测和竞赛评测数据集所代表的环境过于简单，未充分涵盖真实环境下的对抗因素，导致模型在面对真实应用中光照变化、快速运动、相似物体干扰等挑战性因素时适应性较差。此外，评估评测方式均只针对模型进行设计，缺乏对人类视觉系统的评估能力。值得一提的是，图灵测试由于引入了人的评估方式得到相关学者的关注，2015 年布朗大学学者提出 VTT（visual Turing test）测试方案[12]，旨在通过一系列没有歧义的二值问题评估机器是否具有和人类一样的视觉理解能力，虽然尝试通过问答的形式对比机器和人类的能力，但这种评测方式侧重于评估机器对时间、空间和因果关系的综合理解，无法有效度量机器在传统视觉任务上的智能程度与人类的差距（图 2）。

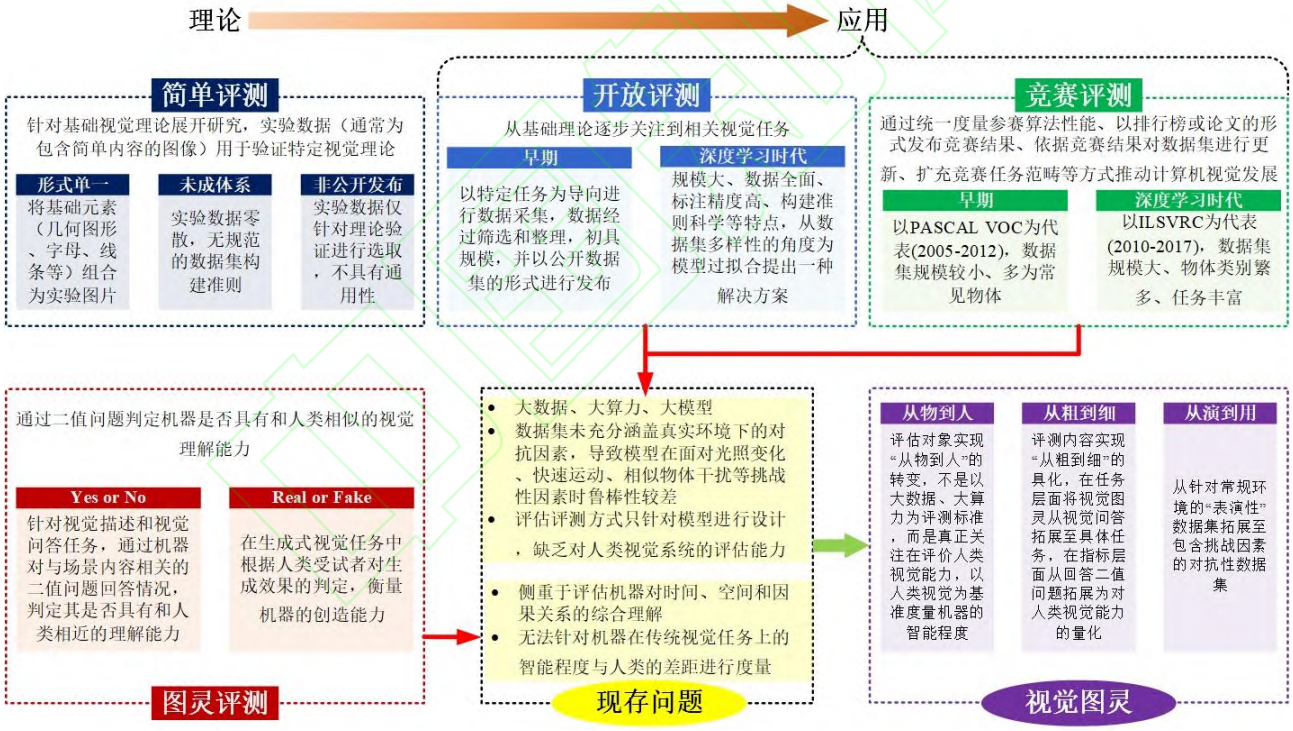

图 2　视觉任务评估评测总结

Fig. 2　The summary of visual tasks evaluation

综上所述，本文从人机对抗评测的角度提出了计算机视觉下一步发展方向：视觉图灵。首先，评估对象实现“从物到人”的转变，不是以大数据、大算力等“外物”为评测标准，而是真正关注在评价“类人”视觉能力，以人类视觉为基准度量机器的智能程度。其次，评测内容实现“从粗到细”的具化，在任务层面将视觉图灵从视觉问答拓展至计算机视觉所研究的具体任务，在指标层面从回答二值问题拓展为对人类视觉能力的量化。最后，评估环境实现“从演到用”的转变，从针对常规环境的“表演性”数据集拓展至包含挑战因素的对抗性数据集。依托于以上3点的突破，计算机视觉技术的发展将不再局限于对大数据和大算力的强烈依赖，而是以人类感知能力为引导，使得计算机视觉研究迈向下一个新的发展阶段，进而为探索实现近似或超越人类的视觉信息感知提供重要的研

究基础。

2 研究现状

如果将计算机视觉理论作为算法的源头，那么评估评测就是算法的落脚。源头决定着算法的天花板，但评估评测决定着算法的应用水平。早期计算机视觉的研究主要集中在对基础视觉理论的完善，提出了MARR视觉理论[13]、成分识别理论[14]等计算机视觉奠基性理论框架，这些理论和框架经过不断发展细化，研究重点逐步聚焦到以物体识别和分类、目标检测和定位、目标分割、目标跟踪等为代表的视觉任务。近几年，大规模数据集为视觉任务提供实验环境，推动了以深度学习为代表的技术发展，并在智慧城市、公共安全、人机交互等应用场景发挥重要作用。

本章从评估评测（评测数据集、评测指标、评估方式）出发，将计算机视觉的发展历程划分为简单评测、开放评测、竞赛评测和图灵评测4个阶段，并对每个阶段的评估评测特点进行梳理和总结。

2.1 简单评测

在计算机视觉发展初期，研究者主要针对基础视觉理论展开研究。此阶段所使用的实验数据（通常为包含简单内容的图像）用于验证特定视觉理论，具有形式单一、未成体系、非公开发布等特点。

1962年，为研究视觉信息的处理机制，神经生理学家HUBEL和WIESEL[15]通过幻灯片对猫展示包含特定模式（具有不同位置和大小的圆形光斑、具有不同朝向和长度的条形光斑）的图像，并记录猫脑神经元在不同模式刺激下的电活动。1963年，ROBERTS[16]通过算法从包含单一几何体的图片中提取简单几何结构，以“积木世界”的方式实现对物体形状和空间关系的描述。1966年，麻省理工学院人工智能实验室的PAPERT[17]举办夏季视觉项目，以包含若干非重叠物体（具有不同纹理和颜色的几何体，如球类、砖块或者圆柱体）的图片为实验对象，尝试通过前景-背景分割完成从图像中自动提取对象。

20 世纪 80 年代，认知科学家以物体识别任务为载体，将基础元素（几何图形、字母、线条等）组合为实验图片，并根据受试者面对不同类型图片的认知表现，对人类视觉认知过程进行解析。1980 年，加州大学伯克利分校 TREISMAN[18]选择由简单元素（不同颜色的字母或不同大小的椭圆）构成的实验图片，对视觉加工问题进行分析，并提出特征整合理论。1982 年，麻省理工学院 MARR[13]提出不同于“积木世界”的视觉计算理论，即通过对心理学、生理学、信息学等领域进行综合，将视觉定义为对外部图像的有效符号描述。1982 年，中科院生物物理所陈霖[19]用正方形、圆形和环形等几何形状组成实验图片，对视觉感知中的拓扑结构展开研究。1987 年，南加州大学 BIEDERMAN[14]在 MARR 视觉计算理论的基础上提出成分识别理论，以包含可拆解物体（如水壶、剪刀、订书机、手电筒、台灯等）的图片为测试数据，分析人类对图片的理解能力。视觉理论的出现标志着计算机视觉成为一门独立学科，并逐步从理论实验向真实应用拓展。

2.2 开放评测

20 世纪 90 年代起，计算机视觉研究从基础理论逐步关注到具体视觉任务。和简单评测阶段的零散实验图片相比，此阶段以特定任务为导向进行数据采集，数据经过筛选和整理，且初具规模，并以公开数据集的形式进行发布。1998 年，LECUN 针对数字手写识别任务发布包含 6 万张 32×32 尺寸图片的 MNIST[8]数据集。2004 年，加州理工学院发布针对目标识别任务的 Caltech-101[20]数据集，其包含 101 类物体、由 9 146 张图像构成，并于 2007 年扩充为包含 256 类物体的 Caltech-256[21]。2009 年，KRIZHEVSKY 和 HINTON[22]发布了针对目标识别任务的 CIFAR 数据集，由 6 万张尺寸为 32×32 的彩色图像构成，具有 CIFAR-10（包含 10 类物体）与 CIFAR-100（包含 100 类物体）2 个版本。

2009 年，斯坦福大学李飞飞[3]教授发布大规模数据集 ImageNet，其在语义学框架 WordNet 的指导下采集包含 2.2 万类物体的 1400 万张图像，为物体识别和分类任务带来全新挑战，标志着计算机视觉进入大规模数据库时代。不同于早期针对特定视觉任务构建的开放数据集，以 ImageNet 为代表的数据集具有规模大、数据全面、标注精度高、构建准则科学等特点，从数据集多样性的角度为模型过拟合提出一种解决方案，并推动了以大数据驱动的深度学习方法的发展。此后，各项视觉任务均出现高质量的代表性数据集，如针对人脸识别任务的 CelebA[23]和 WIDER FACE[24]、针对自动驾驶场景的 KITTI[25]、针对目标跟踪任务的 GOT-10k[26]和 LaSOT[27]、针对场景解析和语义理解任务的 ADE20k[28]和 Cityscapes[29]、针对行人重识别和属性识别的 RAP[30]等数据集。

2.3 竞赛评测

在开放评测的基础上，部分数据集以竞赛的形式发布。竞赛评测通过统一度量参赛算法性能、以排行榜或论文的形式发布竞赛结果、依据竞赛结果对数据集进行更新、扩充竞赛任务范畴等方式推动计算机视觉的发展。

前深度学习时代，竞赛评测以 2005-2012 年举办的 PASCAL VOC[11,31-32]挑战赛为代表。第一届竞赛仅包含 1 578 张图片，针对 4 类物体开展分类和检测竞赛。2007 年，第三届竞赛对数据集规模进行扩充，将物体类别扩大至 20 类，并引入分割和人体部位检测任务。2012 年举办的最后一届竞赛中，数据集规模达到 11 530 张图片，并包含 27 450 个物体标注和 6 929 个分割标注。

2010-2017 年举办的 ImageNet 大型视觉识别挑战赛是近年来计算机视觉领域最具影响力的学术竞赛之一[33-34]，该竞赛从 ImageNet 数据集中抽取部分样本作为竞赛数据，并从最初的图像分类拓展至目标检测、场景分类等任务。2012 年，KRIZHEVSKY 等[35]采用基于卷积神经网络（convolutional neural network, CNN）的 AlexNet 模型夺冠，引发研究者对深度学习方法的关注。此后，GoogLeNet[36]，VGG[37]，ResNet[38]和 DenseNet[39]等模型 ILSVRC 竞赛上展示优异性能，标志着深度神经网络成为视觉任务的主流方法。

除 ILSVRC 之外，VOT[40]，MS COCO[41]，MOT[42]等计算机视觉挑战赛吸引全球科研机构和科技公司参与。VOT[40]是自 2013 年起每年在 ICCV 和 ECCV 研讨会上举办的视觉物体跟踪挑战赛，通过更新评测序列、扩充任务范畴、优化评测指标，实现对复杂环境下单目标跟踪算法性能的评测。MS COCO[41]竞赛起源于微软公司 2014 年标注的同名数据集，图片选取自日常场景，并为每个实例提供额外的分割标注来辅助物体定位。该竞赛以场景理解为目标，包含物体检测、目标分割、人体关键点检测、场景分割等任务。MOT[42]竞赛自 2015 年起针对复杂场景下多目标跟踪任务展开评测，任务场景从 2D 街景下的行人/车辆跟踪拓展至 3D 场景下斑马鱼跟踪，任务范畴从多目标跟踪拓展至多目标分割。

2.4 图灵评测

计算机视觉的发展目标是实现或超越人类视觉感知能力，但简单评测、开放评测和竞赛评测专注于在数据集上算法性能的比较，缺乏与人类视觉能力的对比。在计算机视觉的发展过程中，有学者提出借助图灵在 1950 年提出的模拟游戏思路[43]，以图灵评测的形式对计算机视觉模型开展评估。

已有的视觉图灵评测主要采用视觉描述和视觉问答方式，如通过机器对与场景内容相关的二值问题（Yes/No, Real/Fake）的回答情况，判定其是否具有和人类相近的理解能力；或在生成式视觉任务中根据人类受试者对生成效果的判定，衡量机器的创造能力。

然而，现阶段的视觉图灵工作虽然尝试将人类引入到评测流程中，但其评测形式单一、评测内容宽泛，未有效度量机器的智能程度。以下本文将重点介绍视觉图灵测试，并从视觉图灵出发给出计算机视觉发展的方向。

3 视觉图灵测试

3.1 图灵测试

1950年，英国科学家阿兰·图灵[43]在《计算机器与智能》（Computing Machinery and Intelligence）一文中首先提出了著名的“图灵测试”概念。图灵设计了一个模拟游戏（imitation game），并提出一个问题：“如果游戏中用一台机器代替人类会出现什么情况？”而这也引申出了另一个重要问题，即“机器是否能思考（Can machine think）”？图灵认为，如果询问者无法判断另一个屋子里是人还是机器，那么屋子里的机器就可以称得上是有智能的。

值得一提的是，虽然图灵测试这一概念自诞生以来就引发了广泛而持久的争论[44]，然而图灵测试对于人工智能的重要意义不言而喻，其给出了一种具体可操作的方式来度量智能，即根据对一系列特定问题的反应来决定某一客体是否是智能体。这就为判断智能提供了一个客观标准，从而避免了有关智能本质的无谓争论。比如，从 1990 年开始举办的罗布纳奖竞赛（Loebner Prize Competition）[45]采用标准的图灵测试对机器的能力进行评估。基于图灵测试的人机对抗智能技术也一直是国内外人工智能研究的热点[46]，尤其近年来，以 AlphaGo[47]、冷扑大师[48]等为代表的智能算法在边界确定、规则固定的决策智能问题中已经战胜了人类顶级专业选手，成为图灵测试在智能体评估中的标志性成果。

3.2 视觉图灵研究现状

自 20 世纪 80 年代 MARR 提出视觉计算理论以来，计算机视觉问题也成为人工智能研究的重要组成部分。受相关研究的启发，研究者们开始将图灵测试引入到计算机视觉任务的评估中，并取得了一定进展。其中，最著名的莫过于 2002 年由卡内基梅隆大学提出的 CAPTCHA 测试(Completely Automated Public Turing test to tell Computers and Humans Apart)[49]，也就是俗称的验证码。CAPTCHA 测试通常以文本或图像为载体，使服务器自动产生一个问题并根据相应回答对人类用户和计算机程序进行区分。需要指出的是，CAPTCHA 目的是使人类通过测试而机器无法通过，因此这一技术也被称为反向图灵测试。CAPTCHA 对学术研究和相关技术发展起到了重要的推动作用。目前，CAPTCHA 已经成为一种标

准的网络安全技术，广泛应用于互联网行业。以CAPTCHA为基础，卡内基梅隆大学进一步提出了reCAPTCHA技术[50]来帮助完成典籍的数字化。目前，这一技术已经实现了《纽约时代》报纸扫描存档的数字化。

自深度学习提出以来，计算机视觉在理论和方法上都取得了重要进步。按照经典的机器学习指标，相关算法模型在现有大规模公开评估数据集上已经实现了性能的跨越式提升。那么，如何对机器视觉和人类视觉的能力关系进行有效评估？这些问题受到了研究者们的关注。2015年，布朗大学学者在美国科学院杂志上发表论文，提出了一种针对计算机视觉的图灵（visual Turing test，VTT）测试方法[12]，目的是评估计算机能否像人类一样实现对自然图像的有效理解。在该图灵测试方法中，系统会根据图像的标注内容，按照“故事情节”生成一系列没有歧义的二值问题，而机器和人类可以按照同样的方式进行回答。测试方式如图3所示。

图3 基于视觉问答的视觉图灵测试方案

Fig. 3 Visual Turing test based on visual question answering

基于视觉内容理解的图灵测试也受到了研究者的持续关注。朱松纯等[51]提出了一种针对场景和事件理解的视觉图灵测试。该测试同样采用是非判断的方式，但是测试涉及的场景更为复杂，更加侧重计算机对时间、空间和因果关系的理解能力。除了包含更加复杂的测试场景，有研究者设计了更加复杂的图灵测试问题[52-54]，视觉问答的涵盖范围和回答难度进一步提升，需要围绕计数、物体类别、实例信息等内容进行相应回答。这些研究对于视觉描述（visual caption）和视觉问答(visual question answering, VQA)任务发展起到了积极的意义。

在经典的视觉识别、检测任务之外，越来越多的研究开始关注生成式视觉任务，如图像风格迁移、图像生成和图像渲染等。这类生成任务通常无法采用经典的机器学习指标进行评估，视觉图灵测试成为了评估这类任务效果的一种可行方式。2013年，华盛顿大学和Google的研究者将视觉图灵测试引入到场景重建任务的评估中[55]。在测试中，研究者分别提供一张真实图像和算法渲染后的图像，并要求受试者判断哪一张图像看起来“更真实”。实验结果表明，部分较低分辨率的渲染图像可以通过图灵测试，而高分辨率的图像大概率无法通过测试。作者指出，使低分辨率图像通过图灵测试是三维重建算法短期内可以企及的目标。2015年，麻省理工大学的TENENBAUM等[56]也采用图灵测试的方式对计算机概念学习（Concept Learning）的能力进行评估。TENENBAUM以手写体字符为研究对象，图灵测试的方式与文献[55]较为类似，即同时给出手写体字符和机器生成字符，让受试者判断哪一个字符是由机器产生的。测试结果表明，在手写体字符生成这一任务上机器行为与人类已经很难区分了。文献[57]同样采用了标准图灵测试来对图像染色算法的性能进行评估，测试中32%的算法生成图像成功欺骗了“参与者”。此外，在艺术图像生成效果评估中[58]，研究者在真假判断的基础上还添加了可靠性判断和美感判断的测试内容。可以看出，视觉图灵测试已经成为生成式视觉任务一种重要的评估方式。

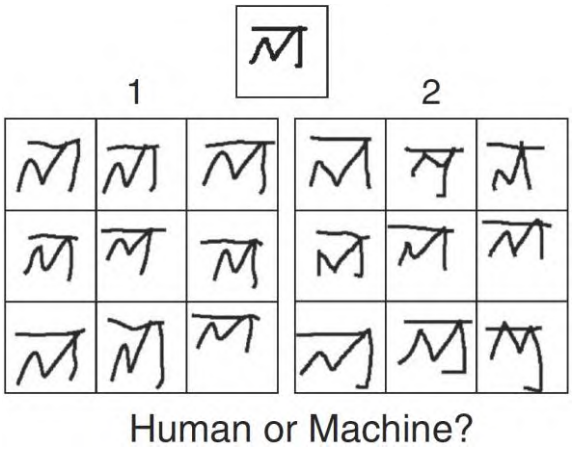

图4 概念学习中的图灵测试[56]

Fig. 4 Visual Turing test in concept learning evaluation

4 展望

随着深度学习技术，海量数据集以及竞赛评测方式的普及，诸多视觉算法在相关数据集上已经达到较高的性能，但当前依赖大算力、大数据的算法在真实应用中表现并不如人意。以人机对抗为评测方式的图灵测试为计算机视觉的发展提出了新的思路。然而，现阶段的视觉图灵工作虽然尝试将人

类引入到评测流程中，但其评测形式单一、评测内容宽泛，未有效度量机器在具体视觉任务上的智能程度，本文从人机对抗出发给出基于视觉图灵的计算机视觉发展可能的方向。

4.1 对象由物到人

正如上述分析，当前计算机视觉关注在数据集的大小，计算资源的多少，称之为“物”，这与计算机视觉是以人类视觉为目标（“人”）的初衷并不符合，而海量标注数据不仅需要大量数据搜集和繁重的标注工作，并且大规模训练对计算资源也提出了非常高的要求。算法性能的提升越来越倚仗算力的堆叠而不是视觉模型和方法的改进。这种研究模式越来越关注数据和算力等“物”的层面，忽略了视觉研究的目的，即机器具备自然（人类）视觉的能力，这种模式对于计算机视觉的发展是不利的。

机器的大规模学习过程与人类的学习过程存在明显的区别。现有最大规模的识别数据集ImageNet 包括约 2 万类物体，其中仅有 1000 类物体图像有较多的标注样本并用于训练和评估。相关研究表明，人类一生可识别的物体种类大约为 3 万类[59]，更重要的是，人类可以在仅获得少量样本的前提下迅速理解新的概念并将其泛化[56],[59]。而目前的评测标准很难从人类学习能力的角度对机器进行更加有效的评估。

尽管现有深度模型在统计学意义的指标上有着优异的表现，但是算法也存在着明显的弱点。文献[60]指出，深度学习模型即使是在识别最常见的物体类别时仍会出现很明显的错误，而人类几乎不太可能出现这类问题。此外，文献[61]发现当给某些图像添加某种程度的噪音时，机器会改变原本给出的高置信度的正确预测结果并做出错误的类别判断，深度模型可以轻易地被对抗样本所“愚弄”。另一方面，相关认知实验[62]表明，人类可以有效辨认对抗样本，而且可以对机器在面对对抗样本时将做出何种判断进行有效预测。这也从一个方面印证了人类的视觉能力要远远超过以大数据大算力为基础的深度学习机器模型。

因此，在接下来的视觉研究中，有必要改变目前基于大数据、大算力的评估标准，将“人”的因素加入到回路中，根据人类的能力对机器的能力进行更加有效的评估[63]。而视觉图灵测试本质上是以类人视觉为标准的一种评估体系，其可以在一定程度上打破机器和人类认知的鸿沟。相信随着对视觉图灵研究的深入，可以使计算机更好的借鉴、模仿人类的视觉和学习过程，从而朝着具备真正意义上的人工智能迈出更踏实的一步。

4.2 任务由粗到细

通过和人对抗来评估智能体能力的图灵测试评估方式越来越得到关注，并取得了一定的进展，对相关领域发展也起到了重要推动作用。但是，正如存在的质疑所提到的，现有的图灵测试方法仍然存在目标不明确、任务宽泛、无法量化等问题，如：①评估所针对的视觉任务相对宽泛；②部分视觉任务缺乏针对性图灵测试设计；③缺乏具体的指标对人类能力进行有效量化等，因此，从粗放式的视觉图灵测试走向精细化的视觉任务测试也是大势所趋。

以视觉问答为例，VTT 涉及物体分类、物体定位和关系推理等多项视觉任务，属于对机器视觉能力的综合考察。因此，很难就机器的某一项具体能力得到可量化的评估结果。而后续针对 VQA 的方法研究[53-54]已经涉及到了视觉与自然语言处理 2 方面的结合，这与最初的视觉图灵测试设定出现了一定的偏差。文献[12]提出，VTT 测试仅仅是一个关于视觉的测试，不涉及自然语言处理的过程（“The interpretation of the questions is unambiguous and does not require any natural language process”）。因此，有必要针对计算机视觉的具体视觉任务进行细化研究。

不同的计算机视觉任务存在着明显的差异，设计一种通用的视觉图灵测试方案较为困难。如，物体跟踪就属于人类视觉中的一项重要能力[64]，视觉问答可以对机器的图像内容理解能力进行评估，但并不适用于直接评估视觉跟踪任务，因为获取、量化人类的视觉跟踪轨迹较为复杂。这就要求研究者根据不同视觉任务的特点进行相应的具体设计。一种可能的解决方案是借鉴视觉显著性的研究过程[65]，采用传感设备对人类的视觉跟踪过程进行有效捕捉，并在此基础上进行视觉图灵测试。

在图灵最初的设想中，如果机器让参与者做出超过 30%的误判，那么可以认为这台机器通过了测试。后续的研究基本按照这个指标对机器的能力进行评估。然而，图灵并没有提出如何对人类的能力进行量化。相关研究表明[66]，人类在不同年龄阶段的视觉认知能力存在明显差异，而现有的视觉图灵测试并没有考虑相关因素。另一方面，在零和博弈任务中就存在对人类能力的具体量化标准，如 Elo 等级分制度，其反映了人类在具体博弈任务上的水平。因此，在视觉图灵测试有必要借鉴相关研究，对机器视觉和人类视觉能力的关系进行可量化的评估。

4.3 数据由演到用

评测数据集是任务评估评测的重要组成。在早期的视觉研究中，视觉理论和框架尚在探索阶段。此时构建的任务大部分是“toy problem”，数据集均较为简单、规模较小，有着明显的“表演”性质。比

如行为识别中的 KTH 数据集和 Weizmann 数据集[67-68]等。这类数据集通常只包含在单一场景下的简单动作。尽管对早期的算法研究和评估起到了推动作用，但是这类数据与真实的应用场景存在明显的差距。

互联网行业的发展，使得海量数据的获取、标注变为可能。而这也推动了以深度学习为标志的大规模训练和评估。此时的数据集类别和样本数量大幅度增加，数据更加接近真实的复杂场景。然而，随着数据规模的进一步提升，数据出现了明显的同质化现象，这并不利于对机器能力的真实评估。此外，统计学的准确率提升并不意味着机器真正具备解决困难问题的能力。

随着计算机视觉从理论走向应用，研究的问题逐渐从简单任务、复杂任务走向对抗任务。比如，某些场景下背景环境会对物体识别带来极大干扰[69]，需要识别的目标存在刻意的隐藏和伪装[70]，篡改伪造内容以混淆视听[71]等。这些对抗因素会对现有的方法带来极大的挑战。因此，有必要设计更加合理的评价体系，对机器在对抗条件下的能力进行更加有效的评估。相比于机器，目前人类仍然具备一定的优势[62]。而人类在对抗视觉任务上的表现可以为机器能力的评估提供重要的参考依据。这也是计算机视觉逼近甚至超过人类的过程中必然要经历的环节。

5 结 论

作为人工智能领域的热点研究方向，计算机视觉已在理论方法、关键技术和实际应用等方面取得巨大进步，但以大数据、大算力为基础的发展模式已无法有效推动计算机视觉下一步发展。本文以算法评估评测（评测数据集、评测指标、评估方式）为主要视角，对计算机视觉的发展历程进行梳理。通过对各阶段存在问题的分析，探讨提出了计算机视觉下一步发展方向：视觉图灵，并提出了3个可能的方向：评估对象实现“从物到人”的拓展、评测内容实现“从粗到细”的具化和评估环境实现“从演到用”的转变，试图推动计算机视觉研究的发展。

总之，计算机视觉的发展推进了人类社会的智能化进程，但依赖大数据、大算力为基础的发展模式和真实场景的需求仍存在差异。视觉图灵为打破现阶段发展瓶颈提供一种可行的思路，为实现近似或超越人类视觉信息感知能力提供重要的研究基础。

参考文献（References）

[1] 黄凯奇, 任伟强, 谭铁牛. 图像物体分类与检测算法综述[J]. 计算机学报, 2014, 37(6):1225-1240.
HUANG K Q, REN Z Q, TAN T N. A review on image object classification and detection[J]. Chinese Journal of Computers, 2014, 37(6):1225-1240 (in Chinese).

[2] 黄凯奇, 陈晓棠, 康运锋, 等. 智能视频监控技术综述[J]. 计算机学报, 2015, 38(6):1093-1118.
HUANG K Q, CHEN X T, KANG Y F, et al. Intelligent Visual Surveillance: a review, 2015, 38(6):1093-1118 (in Chinese).

[3] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, New York: IEEE Press, 2009: 248-255.

[4] https://cs.stanford.edu/people/karpathy/

[5] 中国计算机学会计算机视觉专委会. 未来 5-10 年计算机视觉发展趋势 [EB/OL]. [2021-01-19]. https://www.zhuanzhi.ai/vip/9063e592ca07daedd5e0cd9ba90eb10c.

[6] 胡占义. 计算机视觉简介:历史、现状和发展趋势, 2017. http://vision.ia.ac.cn/zh/teaching/%E8%AE%A1%E7%AE%97%E6%9C%BA%E8%A7%86%E8%A7%89%E8%AE%B2%E4%B9%89%E7%AC%AC%E4%B8%80%E7%AB%A0.pdf

[7] 黄凯奇,谭铁牛. 视觉认知计算模型综述[J].模式识别与人工智能, 2013, 26(10):951-958.

[8] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[9] FERRARI V, JURIE F, SCHMID C. From images to shape models for object detection[J]. International Journal of Computer Vision, 2010, 87(3): 284-303.

[10] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[EB/OL]. [2021-01-28]. https://www.researchgate.net/publication/306218037_Learning_multiple_layers_of_features_from_tiny_images.

[11] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88(2): 303-338.

[12] GEMAN D, GEMAN S, HALLONQUIST N, et al. Visual turing test for computer vision systems[J]. Proceedings of the National Academy of Sciences, 2015, 112(12): 3618-3623.

[13] BARLOW H B. Vision: a computational investigation into the human representation and processing of visual information: DAVID MARR. San Francisco: W. H. Freeman, 1982. pp. Xvi+397[J]. Journal of Mathematical Psychology,1983,27(1): 107-110.

[14] BIEDERMAN I. Recognition-by-components: a theory of human image understanding[J]. Psychological review, 1987, 94(2): 115.

[15] HUBEL D H., WIESEL T N. Receptive fields, binocular interaction

and functional architecture in the cat's visual cortex[J]. The Journal of Physiology, 1962, 160(1), 106–154.

[16] Roberts L G. Machine perception of three-dimensional solids[D]. Cambridge: Massachusetts Institute of Technology, 1963.

[17] MIT Libraries-DSpace@MIT. The summer vision project[EB/OL]. [2020-10-19]. https://dspace.mit.edu/handle/1721.1/6125

[18] TREISMAN A M, GELADE G. A feature-integration theory of attention[J]. Cognitive psychology, 1980, 12(1): 97-136.

[19] CHEN L. Topological structure in visual perception[J]. Science, 1982, 218(4573): 699-700.

[20] FEI-FEI L, FERGUS R, PERONA P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories[C]//2004 Conference on Computer Vision and Pattern Recognition Workshop. NewYork：IEEE Press, 2004: 178-178.

[21] GRIFFIN, G. HOLUB, AD. PERONA, P. Caltech-256 Object Category Dataset[R]. Pasadena: California Institute of Technology, 2007.

[22] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[EB/OL]. [2021-01-28]. https://www.researchgate.net/publication/306218037_Learning_multiple_layers_of_features_from_tiny_images.

[23] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//2015 IEEE International Conference on Computer Vision. NewYork：IEEE Press, 2015: 3730-3738.

[24] YANG S, LUO P, LOY C C, et al. Wider face: A face detection benchmark[C]//2016 IEEE Cnference on Computer Vision and Pattern Recognition. NewYork：IEEE Press, 2016: 5525-5533.

[25] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. NewYork：IEEE Press, 2012: 3354-3361.

[26] HUANG L H, ZHAO X, HUANG K Q. Got-10k: a large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.43(5): 1562-1577.

[27] FAN H, LIN L T, YANG F, et al. Lasot: a high-quality benchmark for large-scale single object tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 5374-5383.

[28] ZHOU B L, ZHAO H, PUIG X, et al. Scene parsing through ade20k dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 633-641.

[29] Cityscapes Dataset. Semantic understanding of urban street scenes [EB/OL]. [2020-12-10]. https://www.cityscapes-dataset.com/

[30] LI D W, ZHANG Z., CHEN X T, et al. A richly annotated pedestrian dataset for person retrieval in real surveillance scenarios[J]. IEEE Transactions on Image Processing, 2019, 28(4): 1575–1590.

[31] HUANG, Y Z., HUANG, K Q., YU, Y N., et al. Salient coding for image classification [C]//2011 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2011:1753–1760.

[32] ZHANG J G, HUANG K Q, YU, Y N., et al. Boosted local structured HOG-LBP for object localization[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2011:1393–1400.

[33] WANG C, REN W Q, HUANG K Q, et al. Weakly supervised object localization with latent category learning.[C]//2014 the European Conference on Computer Vision. New York: IEEE Press, 2014:431–445.

[34] WANG C, HUANG K Q, REN W Q, et al. Large-scale weakly supervised object localization via latent category learning[J]. IEEE Transactions on Image Processing, 2015, 24(4): 1371–1385.

[35] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2021-02-10]. https://arxiv.org/abs/1409.1556.

[36] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2005 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2015: 1-9.

[37] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2021-02-10]. https://arxiv.org/abs/1409.1556.

[38] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2016: 770-778.

[39] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2017: 4700-4708.

[40] Academic and Research Network of Slovenia. Visual object tracking (VOT) [EB/OL]. [2020-12-19]. http://www.votchallenge.net/

[41] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]//2014 European Conference on Computer Vision. Heidelberg: Springer, 2014: 740-755.

[42] Multiple object tracking benchmark[EB/OL]. [2020-11-30]. https://motchallenge.net/

[43] TURING A, HAUGELAND J. Computing machinery and intelligence[M]. Cambridge: MIT Press, 1950.

[44] FRENCH R M. The turing test: the first 50 years[J]. Trends in Cognitive Sciences, 2000, 4(3): 115-122.

[45] SHIEBER S M. Lessons from a restricted Turing test[J]. Communications of the ACM, 1994，37(6): 70-78.

[46] 黄凯奇, 兴军亮, 张俊格,等. 人机对抗智能技术[J]. 中国科学: 信息科学, 2020, 50(4):540-550.
HUANG K Q, XING J L, ZHANG J G, et al. Intelligent technologies of human-computer gaming[J}. SCIENTIA SINICA Informationis，2020, 50(4):540-550 (in Chinese).

[47] SILVER D, SCHRITTWIESER J, SIMONYAN K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[48] BROWN N, SANDHOLM T. Safe and nested subgame solving for imperfect-information games[C]//Proceedings of the 31s International Conference on Neural Information Processing Systems.NewYork：ACMPress， 2017: 689-699.

[49] VON AHN L, BLUM M, HOPPER N J, et al. CAPTCHA: using hard AI problems for security[C]//International Conference on the Theory and Applications of Cryptographic Techniques. Heidelberg: Springer, 2003: 294-311.

[50] VON AHN L, MAURER B, MCMILLEN C, et al. Recaptcha: Human-based character recognition via web security measures[J]. Science, 2008, 321(5895): 1465-1468.

[51] QI H, WU T F, LEE M W, et al. A restricted visual turing test for deep scene and event understanding[EB/OL]. [2020-11-15]. https://arxiv.org/abs/1512.01715v2.

[52] MALINOWSKI M, FRITZ M. A multi-world approach to question answering about real-world scenes based on uncertain input[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York: ACM Press, 2014: 1682-1690.

[53] MALINOWSKI M, ROHRBACH M, FRITZ M. Ask your neurons: a neural-based approach to answering questions about images[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. New York: IEEE Press,2015: 1-9.

[54] GAO H Y, MAO J H, ZHOU J, et al. Are you talking to a machine? dataset and methods for multilingual image question[C]//Advances in Neural Information Processing Systems, 2015, 28: 2296-2304.

[55] SHAN Q, ADAMS R, CURLESS B, et al. The visual turing test for scene reconstruction[C]//2013 International Conference on 3D Vision. NewYork：IEEE Press, 2013: 25-32.

[56] LAKE B M, SALAKHUTDINOV R, TENENBAUM J B. Human-level concept learning through probabilistic program induction[J]. Science, 2015, 350(6266): 1332-1338.

[57] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization[C]//European Conference on Computer Vision. Heidelberg: Springer, 2016: 649-666.

[58] XUE A. End-to-end Chinese landscape painting creation using generative adversarial networks[C]//2021 IEEE/CVF Winter Conference on Applications of Computer Vision. New York: IEEE Press, 2021: 3863-3871.

[59] LAKE B, SALAKHUTDINOV R, GROSS J, et al. One shot learning of simple visual concepts[C]//Proceedings of the annual meeting of the cognitive science society. Merced: University of California, 2011, 33(33).

[60] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//2015 IEEE international conference on computer vision. New York: IEEE Press, 2015: 1026-1034.

[61] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. [2020-10-12]. https://arxiv.org/abs/1412.6572.

[62] ZHOU Z, FIRESTONE C. Humans can decipher adversarial images[J]. Nature Communications, 2019, 10(1): 1-9.

[63] HU B G, DONG W M. A design of human-like robust AI machines in object identification[J]. [2020-10-11]. https://arxiv.org/abs/2101.02327v1.

[64] HYVÄRINEN L, WALTHES R, JACOB N, et al. Current understanding of what infants see[J]. Current ophthalmology reports, 2014, 2(4): 142-149.

[65] AZAM S, GILANI S O, JEON M, et al. A Benchmark of Computational Models of Saliency to Predict Human Fixations in Videos[C]//VISIGRAPP. 2016: 134-142.

[66] SMITH L B, SLONE L K. A developmental approach to machine learning?[J]. Frontiers in Psychology, 2017, 8: 2124.

[67] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]//Proceedings of the 17th International Conference on Pattern Recognition.New York: IEEE Prtess, 2004: 32-36.

[68] GORELICK L, BLANK M, SHECHTMAN E, et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(12): 2247-2253.

[69] HUANG K, WANG L, TAN T, et al. A real-time object detecting and tracking system for outdoor night surveillance[J]. Pattern Recognition, 2008, 41(1): 432-444.

[70] FAN D P, JI G P, SUN G L, et al. Camouflaged object detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2020: 2777-2787.

[71] AGARWAL S, FARID H, FRIED O, et al. Detecting deep-fake videos from phoneme-viseme mismatches[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New York: IEEE Press, 2020: 660-661.