第 50 卷 第 5 期：1842-1854 高电压技术 Vol.50, No.5: 1842-1854
2024 年 5 月 31 日 High Voltage Engineering May 31, 2024
DOI: 10.13336/j.1003-6520.hve.20240404

基于计算机视觉的电力作业人员行为分析研究现状与展望

闫云凤[1,2]，陈　汐[3]，金浩远[1]，齐冬莲[1,2]，储海东[4]，汪金维[5]
（1. 浙江大学电气工程学院，杭州 310027；2. 浙江大学海南研究院，三亚 572025；
3. 香港大学计算机科学学院，香港 999077；4. 国家电网有限公司，北京 100031；
5. 浙江天衡五维电子科技有限公司，天衡 310027）

摘　要：电力作业人员的有效监管是保障电力安全生产的基础。该文对电力视频中作业人员的行为识别研究进行了归类总结，涵盖静态行为分析(穿戴分析、动作分析和组合分析)和动态行为分析(复杂动作、时序行为和行为预测等)；详细综述了电力作业行为分析中的核心算法模块，包括目标检测、姿态估计和视频跟踪等；论述了电力作业行为识别在算法高效性、鲁棒性、灵活性等方面所面临的应用难点和挑战，并展望了电力作业行为智能监控领域的未来发展方向，特别强调了在软硬件结合、通用大模型、生成式人工智能方面进行技术创新和改进所蕴含的潜在机会。

关键词：行为分析；视觉理解；电力监控；目标检测；姿态估计；视频跟踪；行为预测

Research Status and Development of Computer-vision-based Power Workers' Behavior Analysis

YAN Yunfeng[1,2], CHEN Xi[3], JIN Haoyuan[1], QI Donglian[1,2], CHU Haidong[4], WANG Jinwei[5]
(1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China; 2. Hainan Institute of Zhejiang University, Sanya 572025, China; 3. Department of Computer Science, The University of Hong Kong, Hong Kong 999077, China; 4. State Grid Corporation of China, Beijing 100031, China; 5. Zhejiang Tianheng Wu Wei Electronics Technology Co., Ltd., Tianheng 310027, China)

Abstract：Effective supervision of power workers is the basis for ensuring the safe power production. In this paper, we categorize and summarize the research on the behavior recognition of power workers in video, including static behavior analysis (clothing analysis, action analysis, and combination analysis) and dynamic behavior analysis (complex actions, temporal behavior, and behavior prediction, etc.). We provide a detailed overview of the core algorithmic modules in the analysis of power operation behaviors, including target detection, pose estimation, and video tracking, and others. We also discuss the challenges and difficulties in applying these techniques in terms of algorithm efficiency, robustness, and flexibility, and put forward the future development direction of the field of intelligent monitoring of power operation behaviors; meanwhile, the emphasis is put on the potential opportunities in technological innovation and improvement through the integration of hardware and software, general large models, and generative artificial intelligence.

Key words：behavior analysis; visual recognition; power monitoring; object detection; pose estimation; video tracking; behavior prediction

0　引言

电力系统在社会生产生活中扮演着至关重要的角色，而电力作业人员行为的准确分析和及时预警是保障电力系统安全运行的关键。作业人员行为的精准识别能够有效地消除生产过程中的安全隐患，确保电力系统在复杂多变的环境中稳定地运行。

随着计算机视觉技术的迅猛发展，目标检测、人体姿态估计、多目标跟踪、视频理解与预测、多模态大语言模型等技术推动了电力作业人员行为识别向数字化、智能化发展。目标检测技术，作为计算机视觉领域的基石，已经在电力行业中发挥了重要作用。通过精准地定位和识别图像中的安全帽、

基金资助项目：国家自然科学基金(62127803)；浙江省科技计划项目(2022C01056)。
Project supported by National Natural Science Foundation of China (62127803), Key R&D Project of Zhejiang Province (2022C01056).

安全带等关键物体，该技术能够有效地监测电力作业人员是否遵守安全规范，从而预防潜在的安全事故[1-2]。人体姿态估计技术则通过识别人体的关键点，分析作业人员的动作和姿态，进一步判断是否存在违规操作或危险行为[3]。这项技术的应用，使得电力监控系统能够更加细致和深入地理解作业现场的动态情况，为安全管理提供了新的视角。目标跟踪技术在电力场景中的应用，使得系统能够同时跟踪多个作业人员，记录他们的行为轨迹和相互之间的交互。这对于分析团队协作、预防群体性违章行为具有重要意义，同时也为事故后的追溯和分析提供了宝贵数据。而视频理解与预测技术的发展，让电力监控系统不再局限于对当前行为的分析，而是能够通过分析视频序列中的时空关系，预测作业人员的未来行为[4]。这种前瞻性的分析为电力安全管理提供了更为智能的手段，有助于提前发现并防范潜在风险。多模态大语言模型的引入，为电力视觉技术带来了新的变革。这些模型结合视觉信息和语言描述，能够提供更为丰富和灵活的交互方式，使得非专业用户也能够方便地与电力监控系统进行交流，获取所需的信息和支持。例如：通过自然语言查询，用户可以快速了解作业现场的实时情况，或者询问关于特定安全规程的问题。

本文详细综述了电力监控中人员行为识别和预警方法的研究现状，并对各算法进行功能模块化拆解;介绍了电力视觉算法的基础原理和最新发展，主要包括：目标检测、人体姿态估计、多目标跟踪、视频理解与预测、多模态大语言模型等；通过详细的算法分析，进一步探讨了如何将最先进的方法应用于电力领域，并讨论了其中可能需要进行的改进与适配；最后，深入分析了电力作业人员行为识别任务所面临的挑战，从而为未来的研究方向和技术发展提供有益的参考。

1　电力作业人员行为分析任务解析

电力作业行为呈现背景复杂、形态多样、遮挡密集、行为时序交互关联、人物大量交互、人工经验干预、速度精度要求高等特点，相比于常见的行为分析任务，电力作业行为识别具有以下特殊性：(1) 电力视频采集环境恶劣，电力作业现场摄像机架设位置往往比较局限，常为固定机位下的大场景视角，使得作业过程视频产生帧率低、画质差、相机远以及由于电力设备密集出现人员遮挡等情况，导致识别难度增大；(2) 电力作业多为带电操作，与设备大量交互，点多面广、工具复杂，被检测目标的外观、姿态、角度及操作设备等在作业过程中不断变化,给目标检测与行为识别带来了很大挑战；(3) 在电力运维时，作业人员一般穿戴类似工作服，在作业时互相交互、遮挡，动作变化幅度大，拍摄视角随着人员的移动也会发生较大变化，给目标跟踪任务带来极大人员重识别难度，对算法鲁棒性提出了较高的要求；(4) 作业违规行为隐患大，且事故传播快，受限于现场部署的通信压力、计算资源及数据保密等客观条件,要求识别方法快速且准确。

从应用层面，电力作业人员行为识别任务大体可分为静态行为分析和时序行为分析两大类，如图 1 前半部分所示。静态行为分析是指依靠单帧图像即可完成的行为分析，包括违规穿戴检测、违规动作分析、组合行为分析等。静态分析主要依赖于 2D 图像处理算法，如目标检测、图像分类、人体姿态估计、人-物交互检测等。时序行为分析需要对事件的先后发生关系进行考量，如工序监测、人员轨迹分析、人员未来行为预测等。时序分析涉及视频处理算法如视频理解、视频跟踪、时空预测、多模态理解等。在监控过程中，智能系统从视频流中以一定的帧率提取图像，依照不同要求，对图像帧进行单独处理或对多帧图像进行统一处理。图 1 后半部分展示了电力监控场景中各类行为分析任务示意图，通过计算机视觉模型和算法的相互配合，可以解决电力生产环境中的多种复杂的行为分析问题。

本章将自顶向下对电力作业人员行为识别任务进行解析。主要探讨计算机视觉可以解决的具体任务、对复杂的监控系统进行结构化拆解，并进一步探讨算法的具体应用情况。

1.1　静态行为分析

电力作业行为分析的大部分任务可以通过图像分析的手段解决。静态行为分析包括穿戴分析、动作分析和组合分析 3 个大类。

1）穿戴分析

穿戴分析是电力作业行为识别的首要环节之一。通过计算机视觉算法，系统能够对工作人员的外观属性进行全方位监测，并且返回标签。其中，目标物体检测技术可用于检测是否戴有安全帽、是否绑定安全绳及是否按规定穿戴工装等。

这些信息的准确获取有助于预防潜在的安全隐患，确保工作人员的作业安全。穿戴分析主要依

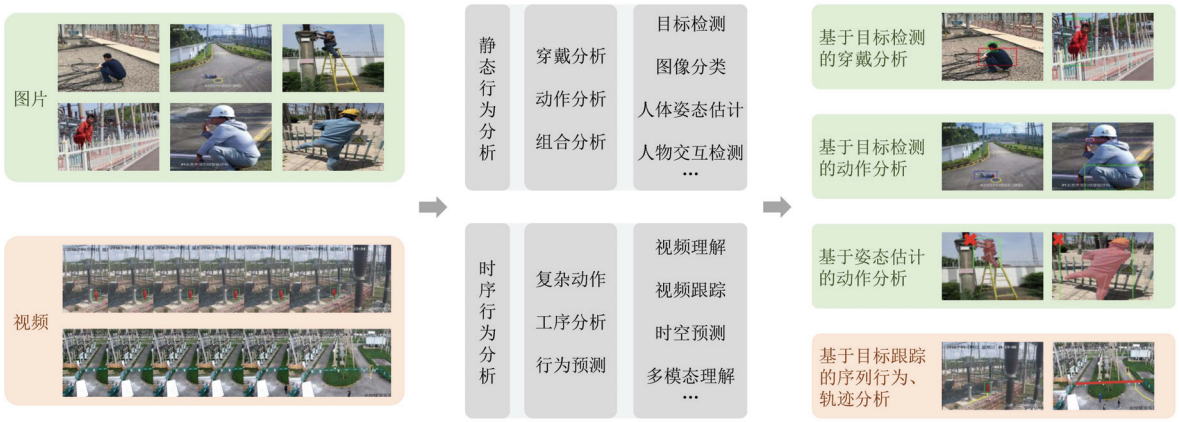


图 1 基于计算机视觉的电力安全行为分析任务拆解及算法

Fig.1 Task decomposition and algorithms for power safety behavior analysis tasks based on computer vision

赖于计算机视觉中的目标检测、目标分类算法参与。例如：利用卷积神经网络进行目标分类，或利用目标检测模型对违规穿戴进行坐标定位，有效解决了电力监控中的诸多检测难题[5-7]。

2）动作分析

动作分析是电力行为识别中的另一关键步骤。通过人体动作分析与姿态估计技术，能够准确地捕捉到工作人员的各种动作，从而判断是否存在违规行为或危险动作。例如：识别工作人员是否采取了不安全的动作姿势或者是否进行了不符合规定的操作，包括在禁烟场所抽烟，伸手触碰带电设备，在违规区域躺卧等。动作分析算法主要依赖于人体姿态估计、目标检测、目标分类算法的协作配合。以人体姿态估计模型为基础，对电力监控中的多种违规行为进行预警与监测，如抽烟、翻越围栏、违规触碰等[8-10]。

3）组合分析

除以上单体分析外，还有很多较为复杂的电力违规行为不能仅仅依靠单个人员的行为进行判断。例如：对工具是否违规使用需要人-物交互的检测和识别；是否闯入禁行区域需要对人员位置以及禁行区域同时进行分析；一些复杂的规章如“人员使用梯子无人扶梯”，需要对多个人员的动作以及对工具的交互进行联合识别；“聚众、打架”等行为需要对多人之间的交互动作进行分析。

区别于单体分析，此类组合分析需要多个子模块算法的相互合作，串联成更加智能的系统。例如：使用目标检测模型与人体姿态估计模型相互配合，对人员手持绝缘子的违规姿势进行检测[11]；组合目标检测、图像分类等多种模型，对变电站监控场景进行安全监控[9]；利用人体姿态估计定位人员手部、头部等区域，从而分别进行违规动作、穿戴检测，如抽烟、安全帽检测等[12]；结合人体姿态模型和注意力机制，对不同人体区域的违规穿戴进行联合检测与识别[2]。

1.2 时序行为分析

1）复杂动作

虽然大多数具体动作可以依靠单帧静态图像进行识别，但较为复杂的动作则需要依靠视频帧之间的连续性进行识别。例如：电力维修人员双人扛梯奔跑时，需要通过视频帧之间的连续性和动作的时序信息来准确识别这一紧急响应行为，否则难以和静止状态或缓慢行走进行区分[13-15]。

2）工序分析

工序分析需要考察人员行为的前后依存关系。其核心难点是对于某一个固定人员的位置和行为进行跟踪。此外，在电力行业中，比如在进行倒闸操作时，不仅需要精确跟踪操作人员的位置和行为，还需要分析他们的行为模式和操作顺序，确保每一步操作都符合安全规程，否则很难对此类时序动作进行判断[16-18]。

3）行为预测

作为电力作业人员行为识别的前瞻性任务，借助目标跟踪和时空预测技术，系统可以分析工作人员的轨迹，并预测未来可能发生的违规行为。这有助于提前采取措施，防范潜在的安全风险，为电力系统安全管理提供更为智能的手段[19-20]。

2　电力作业行为分析相关算法

电力作业行为分析是极其富有挑战性的课题，一个完备的电力行为识别系统往往有多个不同的算法模块构成。前文从系统层面对电力行为分析中的具体任务进行总结，本章将对电力作业行为分析领域相关的视觉算法进行探讨。

2.1　目标检测

目标检测算法在电力作业行为分析中起到最基础的作用，可以直接对静态违规行为、穿戴、设备、异物进行定位和分类，也可以对人员位置进行预先定位，从而作为后续算法的输入。

目标检测算法在近年来取得了快速进展，从最初以卷积神经网络 Faster-RCNN[21]为代表的两阶段目标检测，过渡到以 YOLO 系列[22]为代表的一阶段框架，到视觉 Transformer[23]时代提出的以 DETR[24]为代表的端到端目标检测架构。目标检测算法最前沿方向是融入图文多模态模型 CLIP[25]的开放词汇目标检测。目标检测算法在精度和广度方面逐步提升，力求为真实世界复杂场景理解提供助力。

1）基于卷积网络的目标检测

在深度学习发展的初始阶段，目标检测框架主要以卷积网络为基座。以 Faster-RCNN 系列算法为代表的两阶段框架是目标检测的主流[26-27]。该框架第一阶段使用特征提议网络(RPN)对存在检测目标的区域进行初步判定；第二阶段对初始提议框进行坐标精确回归和类别预测。在特征金字塔(FPN)[28]、兴趣区域对齐(ROI Align)[29]等技术的辅助下，新兴的两阶段算法在各类数据基准上均取得了较高的精度[30-32]。但两阶段算法结构复杂，速度较慢。为解决两阶段模型的速度瓶颈，以 YOLO 系列为代表的一阶段检测架构受到了工业界的广泛关注[33-36]。一阶段算法不需要特征提议网络参与，大多在锚点的基准上直接预测目标物体的位置和类别，从物体的中心点出发对物体边界框进行回归[37-39]。在电力行业中，一阶段方法也被成功应用于行为检测。例如：通过使用 SSD(Single Shot MultiBox Detector)框架，研究者能够有效地检测和识别电力系统中的操作行为，如开关设备的倒闸操作[40-41]。这种实时的行为检测对于提高电力系统的安全性和操作效率具有重要意义。近年来，研究者对于一阶段模型的主干架构、激活函数、样本增强策略和损失函数进行了深入研究，其在精度上已可对齐甚至超过经典的二阶段模型，且在运行速度和效率上具有显著优势。

尽管如此，基于卷积网络的目标检测算法仍然存在一些弊端。例如：经典的卷积模型主干(如 ResNet)的容量有限，无法很好地利用超大规模数据集的信息；基于卷积模型的目标检测需要非极大值抑制(NMS)作为后处理，对于模型的泛化性和运行速度仍然存在影响。

2）基于视觉变换器的目标检测

视觉变换器(Vision Transformer[23])的提出，显著提升了各类视觉感知任务的效果。Transformer 架构移除了卷积神经网络的归纳先验，大幅提升了模型的容量，在大规模数据预训练/自监督训练的加持下，模型特征的表达能力也得到显著提升[42-43]。与此同时，以 DETR 为代表的基于 Transformer 的目标检测架构首次去除了复杂耗时的 NMS，真正意义上实现了“端到端”的目标检测[44-45]。

基于 Transformer 的目标检测架构首先使用 Transformer 视觉编码器将完整图像分块进行特征提取，编码器中的注意力机制使得模型对图像中的长距离依赖关系进行更好的建模。Transformer 解码器使用可学习的查询令牌(Learnable Query Token)表示图像中的目标物体，与图像特征进行多轮交互后，获得目标物体的准确表征。在此表征基础上，回归出各个目标物体的坐标和类别。基于视觉变换器的目标检测算法大幅提升了检测精度，且为视觉模型参数和数据量的大规模扩展提供了可能。在电力行业，Transformer 架构也被应用于行为检测任务中。例如：研究者利用 Transformer 模型对电力设施维护人员的行为进行精确识别和分类。通过分析工作人员与电力设备之间的交互行为，Transformer 模型能够有效地区分不同的操作流程[46-47]，如更换保险丝、检修电路板等，从而提高电力系统的安全性和维护效率。

3）基于多模态对齐的目标检测

CLIP 模型的提出，打通了视觉和语言之间的模态壁垒，大幅释放了视觉模型对于开放世界的理解能力[25]。之前的目标检测算法仅能对预选训练的类别进行检测，而真实世界的事件非常复杂，往往无法使用少数预定义的标签涵盖。为此，CLIP 提出使用“视觉-语言对比学习”的方式将图像特征与相对应的语言描述特征映射到相同的空间，将视觉分类任务重离散化的“标签”替代为灵活的“语言描述”。

视觉-语言对齐的提出为各个视觉感知算法提

供了新的研究思路。目标检测算法也随之步入了开放词汇时代，以 GLIP[48]和 Grounding-DINO[49]为代表的开放词汇检测算法在 CLIP 的基础上更进一步，将图像-语言的对齐训练转换为区域-语言的对齐。利用 CLIP 强大的先验知识和图文理解能力，现有的开放词汇检测算法可检测任意文字描述对应的目标，该目标可以是物体类别、动作描述、外观描述等，也是较复杂的物体组合、物体局部区域[50-51]。如图 2 所示，浙江大学研究的 PowerGPT 使用 CLIP 和一个线性层将电力图像视觉信息映射到词嵌入空间，通过语言模型生成问答结果。

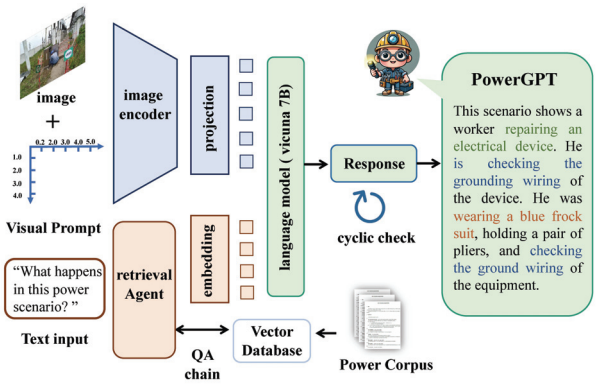


图 2 PowerGPT 视觉问答

Fig.2 PowerGPT visual questions and answers

2.2 人体姿态估计

人体姿态估计是电力作业行为分析的关键。通过姿态估计可以直接对某些违规、危险行为进行识别，或者通过人体关键点的位置对相关区域(如手部、脚底、头部)进行着重分析。

人体姿态估计也是计算机视觉的基础问题，从对单人图像进行处理，到高速同步处理多人图像，再到 3D 空间内对人体进行准确参数化建模，人体姿态估计建模的精准程度和可用性逐步提升，如图 3 所示。

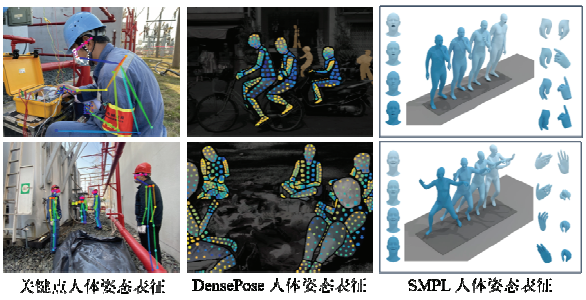


图 3 人体姿态估计表征发展历程

Fig.3 Evolution of representation for human pose estimation

1）单人姿态估计

单人姿态估计是人体姿态估计的基础子任务，输入给定人员图像，要求模型定位人体主要关节点在 2D 图像中的位置坐标和类别。单人姿态估计以自底向上的框架为主。在预训练的视觉主干模型上，增加多尺度特征融合模块和对应的关键点检测头，预测对应关键点的热力图，随后从热力图的峰值回归出对应的关键点坐标。单人姿态估计任务的关键在于如何在扩大模型感受野的同时，保留模型对于局部细节的敏感性。由此，需着重探索设计多尺度融合模块，从而提升模型的特征表达能力。在电力行业中，单人姿态估计的应用场景非常广泛。例如：在监控电力设施维修人员的操作过程中，通过单人姿态估计技术，可以实时监测工作人员是否正确使用工具，是否采取了正确的安全站立姿势等。这对于提前发现潜在的安全隐患、预防事故发生具有重要意义。在扩大模型感受野的同时，保留模型对于局部细节的敏感性是提高单人姿态估计精度的关键。特别是在复杂的电力环境中，工作人员可能处于各种姿态，且背景复杂多变，这就要求模型能够准确地捕捉到关键点的细微变化。因此，设计有效的多尺度融合模块，提升模型的特征表达能力，对于提高单人姿态估计的准确性和鲁棒性至关重要。通过这种方式，可以更加精确地识别和分析电力工作人员的动作，为电力行业的安全生产提供强有力的技术支持。

CPM[52]和 Hourglass[53]模型作为单人姿态估计的代表性方法，通过多阶段堆叠特征提取模块进行更好的特征建模。HRNet 设计了精巧的多尺度特征融合策略，使得模型在保证细节分辨率的同时获取全局信息[54]。随着视觉 Transformer 主干的大规模应用，ViTPose 使用视觉 Transformer 模型作为主干，可提取更丰富的多尺度特征，从而达到更好的效果[55]。

2）多人姿态估计

早期的多人姿态估计以自顶向下(Top-down)的框架为主，即先用目标检测模型检测出每一个人员的位置框，随后在每一个目标框内进行单人姿态估计。代表性方法为 RMPE[56]和 Mask-RCNN[29]。但自顶向下的方法需迭代处理每一个目标框，计算效率较低，并且无法很好地处理较复杂的重叠场景。为应对该挑战，多种自底向上(Bottom-up)的方法被提出，大体分为关键点检测和关键点聚类。第一阶段检测出图像中出现的所有人物的关键点；第二阶

段对关键点属于的具体人员进行归类。PersonLab 预测关键点之前的偏移量(offset)对不同的关键点进行链接[57]。OpenPose 提出局部关联场(Part Affinity Fields)实现快速的关节点连接[58]。类似方式也不断被提出，通过预测关联场或对关键点的特征进行聚类，从而实现多人姿态估计[59-60]。Transformer 的提出为多人姿态估计提供了新思路，类似于端到端目标检测的 DETR 架构，可采用 Transformer 的查询表示代表每一个独立的人员，在此基础上进一步回归每一个关键点的坐标与类别[61]。

3）3D 姿态估计

2D 姿态估计是在图像的像素空间回归人体主要关节点的坐标。该方法可以应对大多数场景下对于人体的建模需求，但无法对人员的具体身形、细节动作、空间坐标进行良好表示。为解决这一问题，研究者们提出用更加细致的 3D 表征对人体姿态进行估计。

3D 人体姿态估计的主流方法分为两种：一是参数化人体模型，采用事先定义好的人体模型(如 SMPL[62])来表示 3D 形状，网络只需要得出模型的参数，然后利用参数进行可视化。二是预测得到更多的关键点，用密集的关键点表征得到类似 3D 重建的效果，然后进行网格化的可视化表示，代表方法为 DensePose[63]。目前广泛采用的是基于 SMPL 的变种，该模型定义人体模板网格，包含 6890 个顶点和 13 776 个面片。该模型使用 10 维参数向量来控制人体形状，利用 24 个关节点旋转参数来掌控人体姿势。每个关节点旋转参数都以 3 维向量的形式表示，分别表示该关节相对于其父关节沿 x、y、z 轴的旋转角度。进一步，SMPL-X 模型采用更多的顶点进行精细建模，同时引入了参数化控制面部表情和手部姿态[64]。这两个模型提供了规范的、通用的人体参数化表示，并与工业 3D 软件(如 Maya 和 Unity)兼容，提出一套简单而有效的蒙皮策略，确保了人体表面的顶点在关节旋转运动时不会出现明显瑕疵。近年来，还涌现了一些改进的人体模型，如 SoftSMPL[65]、STAR[66]、BLSM[67]、GHUM[68]等。

2.3　视频目标跟踪

视频目标跟踪算法对于人员时序行为分析尤为重要，如图 4 所示。视频目标跟踪算法可以检测视频序列中每一个人员的位置坐标，并且绑定同一人员在不同视频帧的对应关系，从而达到对人员长期行为、位置时序关系的建模与分析，为人员未来的轨迹预测提供可能。视频目标跟踪任务分为单目标和多目标跟踪，电力视频目标跟踪主要是多目标跟踪。多目标跟踪算法的难点在于解决视频运动中的遮挡、变形、运动模糊等问题，并且对于离开画面的人员在重新进入跟踪画面后正确匹配之前的跟踪序列。

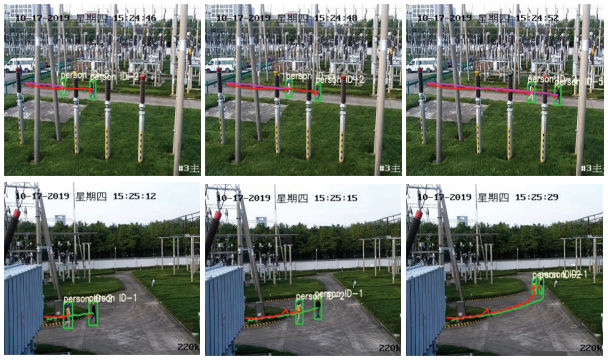


图 4　多目标人员跟踪示意图

Fig.4　Demonstrations for multi-object human tracking

1）基于目标检测的跟踪

经典的多目标跟踪方案将目标检测和跟踪匹配作为两个独立的阶段。该经典框架被称为 Tracking-by-detection。具体而言，该框架先对视频逐帧进行目标检测，获得所有人员的具体位置框，然后对每一个人员区域框提取对应的深度表征，从而使用表征相似度、框之间的 IOU 重合度、通过卡尔曼滤波预测的未来位置等信息，对不同帧中检测出的人员框进行匹配，代表性方法为 SORT[69]、Deep-SORT[70]、ByteTrack[71]等。先检测后跟踪的框架优点是简单有效，但其跟踪性能受到了第一阶段检测算法的制约。

2）目标检测-跟踪联合学习

基于联合学习的范式将目标检测和跟踪合并到同一个阶段，使用 FPN 结构混合每一帧图像的多尺度信息，在不同尺度的特征图上分别接入跟踪预测头，对每一个目标的位置框和匹配特征进行同时预测。之后，使用特征匹配的方法将每一帧独立的坐标框连接成为跟踪序列。JDE[72]作为代表性方法，在实时的速率上实现了稳健的跟踪效果。其他方法(如 FairMOT[73])对目标检测和人员特征提取做出平衡，取得了更高的性能。CenterTrack 将人员表示为点，根据中心区域的特征，对人员的位置和匹配关系进行推断[74]。

3）端到端多目标跟踪

Vision Transformer 提出以来，类似于上文介绍

的端到端目标检测和姿态估计，多目标跟踪算法也进化出了端到端的模型。TransTrack[75]、TrackFormer[76]、ContrasTR[77]用多个可学习查询令牌表示待跟踪的人员，在 Transformer 编码器/解码器提取特征后，在令牌上对人员的位置坐标和匹配特征进行预测。

2.4 视频理解与预测

电力系统是一个高度动态且复杂的环境，其中涉及到大量的设备操作和人员交互。为更深入地理解和预测电力场景中的行为，视频理解技术必须超越传统的单帧图像分析，转而关注视频中的时序信息和时空关系。计算机视觉中的“视频理解”任务力求对视频中依赖时序关系的行为进行识别。与此同时，利用视频帧间一致的特性以及对视频内容的理解，“时空预测”任务可以利用视频的前序帧对尚未发生的时间进行推断。

1）视频理解

该任务将一整段视频直接输入模型进行标签预测。在电力系统中，视频理解技术可以用于监测和分析设备状态、操作流程及人员行为。例如：通过分析变压器冷却风扇的旋转速度，可预测设备可能的过热情况；或者通过监测操作人员的行动，确保安全规程的遵守。初始时期的视频理解使用双流网络分别对空间和时序信息进行建模[78]，或者使用 3D 卷积直接进行时空特征提取[79-80]，这两种方法尝试更好地结合空间和时序信息，对电力操作流程进行建模，以识别不规范的操作行为。TSN 将视频切分成小片段进行特征提取，提升了视频理解的效果和运行效率[81]。Non-local 在卷积中加入长程注意力机制[82]，大幅增强模型对视频中长距离依赖关系的理解，以识别人员是否穿戴了必要的安全装备或是否遵循了安全操作规程。此外，视频理解技术还可以用于监测高压输电线路，通过识别导线舞动等现象，预防输电线路事故的发生。SlowFast 作为视频理解的经典工作，以不同的采样率分别提取视频的空间细节和时序事件[83]。Transformer 时代以来，TimeSforme[84]、VIViT[85]、ViDTr[86]、MViT[87]等直接将 Vision Transformer 扩展到视频领域。

2）时空预测

该任务主要是给定初始序列对后续序列进行预测，该序列可以是传统的视频帧或者车流、人流的实时统计数据等。在电力系统中，时空预测技术可以用于预测电网负载、设备故障概率以及人员行为趋势。时空预测模型大体可分为基于循环神经网络的方法和单步预测的方法。ConvLSTM[88]第 1 次将卷积神经网络和长短期记忆网络相结合。在此基础上，一系列方法对记忆模块、训练模式、基础模型设计等方面做出改进，提升时空预测模型的性能，其改进版本 ConvLSTM 利用历史电网负载数据和实时视频监控数据，预测未来的电网负载变化，以优化电力分配[89-90]。然而，基于循环神经网络的方法有计算量大、结构复杂等缺点。与此同时，基于单步预测的方法(如 SimVP[91]、TAU[92]、DMVFN[93]等)，通过精巧的注意力结构设计和运动信息捕捉，在大大优化模型计算效率的同时，达到匹配基于神经网络模型的效果。通过单步预测方法实时进行预测，这些方法可结合设备操作视频和历史故障数据，预测设备可能出现的故障，提前进行维护。此外，研究基于高效注意力机制的模型，可在高效预测的同时更准确地分析人员在电力场景中的行为模式，预测其未来的行为趋势，以提高安全管理的效率。

2.5 多模态大语言模型

浙江大学齐冬莲教授团队的 PowerGPT 将开源大语言模型迁移到电力业务场景中，实现电力人员行为分析、缺陷检测、业务知识问答等任务。2022 年，ChatGPT 的提出展示了大规模自回归语言模型对于真实世界的强大理解能力，其用语言对话的方式解答用户几乎关于任意话题的提问。2023 年，GPT4V 更是扩充了大语言模型对于图像的理解能力。由此，视觉感知算法的人-机交互模式步入了新的时代，从以往需要预先定义的离散标签转变成更容易被用户接受、泛化能力更强的视觉问答。多模态大语言模型以任意的图像和视频作为输入，无需进行额外的训练即可根据用户提问对图像和视频中发生的事件进行解读。

大语言模型是基于 Transformer 架构，以预测下一个单词为基本学习形式，在巨大的参数量(十亿到百亿)和训练数据量(几乎互联网可获取的所有文本)加持下，涌现出超越已有人工智能模型的理解与泛化能力。随着大语言模型 LLaMA[94]的开源，研究者快速致力于大语言模型的迭代以及提升其多模态理解能力。LLaVa[95]作为代表性方法，使用 Image Caption 模型生成的视觉问答对大语言模型进行微调，从而使大语言模型具备进行视觉问答的能力。Mini-GPT4[96]提出仅需微调图像编码器的映射层，即可将图像表征融合进入语言表征进行语言问答。

Qwen[97]开源了一系列具有媲美 GPT4 能力的多模态大语言模型，给社区的技术迭代提供了极大的便利。GPT4ROI[98]、Kosmos[99]、Shikra[100]等更进一步通过用户指定区域进行交互问答，支持点、框等多种格式的视觉指代，由此支持大语言模型对图像或视频进行更加精确细致的理解。与此同时，大语言模型的视觉理解与问答能力逐步被拓展到视频，支持视频或电影中复杂事件的推理与问答[101-103]。浙江大学齐冬莲教授团队基于 22 万幅电力图像构建出 80 万条高质量电力问答指令对话库，包含自然场景、红外、航拍、文本、表格等多种数据模态，基于 LLAVA V1.5-13B 构建出电力视觉问答大模型 PowerGPT，涵盖输电、变电、运维等电力业务场景。如图 5 所示，如今的多模态大语言模型已经具备理解复杂图像场景和视频序列中发生的细节事件以及因果关系的能力。

在大模型中，BERT 和 CLIP 是两大关键技术。BERT 的预训练任务是通过掩码语言模型和下一段预测两个阶段进行的。在掩码语言模型中，输入文本的一部分被随机掩码，模型需要预测这些被掩码的词。而 CLIP 的创新之处在于，它能够将图像和文本映射到一个共享的向量空间中，从而使得模型能够理解图像和文本之间的语义关系。BERT 和 CLIP 的技术给大模型提供了高效的训练流程，极大的推动了大模型的发展。

目前，大模型的技术方案分为 4 个阶段：提示工程、AI 智能体、大模型微调、预训练。其中最基础的即预训练技术，通过构建深度学习网络框架，大规模数据处理，将大量数据送入模型训练。在 GPT 的迭代过程中，GPT-1 最初具有 117M 模型参数，共有 12 层 Transformer 层数，使用 Books1 和英语维基百科数据进行训练；到 GPT-2，模型规模扩大到 1.5B，Transformer 层数增加到 48 层，训练数据加入了 Books2；到 GPT-3，模型规模进一步扩大到 175B，Transformer 层数增加到 96 层，训练数据加入 Common Crawl 和 WebText2，而 GPT-3 的效果也成功引起了全世界的轰动。预训练技术需要消耗资源，但能达到最好的效果。第二个常用的技术是大模型微调技术，由于预训练成本高，且基础模型缺少对特定领域的数据理解，因此许多个性化服务都采用私有化的大模型微调技术。大模型微调通过在原始模型中加入少量可学习的 Adapter 模块，而不是训练模型的所有参数，并通过残差连接和跳



图 5　多模态大语言模型效果展示

Fig.5　Demonstrations for the abilities of multi-model large language model

跃连接结构提升微调的稳定性，从而以很低的成本学习特定领域数据中的特征。大模型微调技术具有很强的适用性，对于基础大模型私有化领域应用具有很高的性价比。第三个技术是提示工程技术，大模型具有上下文学习能力，对于同一个问题，直白的提问、给出一个思路或将问题拆解，都会使大模型产生不同质量的回答，通过提示工程，自动根据用户提供的输入生成提示，从而使大模型发挥出最有效的能力。提示工程是所有方法中门槛最低、最容易产生效果的，但其并没有对大模型本身有本质的改变。最后一个技术是 AI 智能体技术，该技术旨在赋予大模型规划、记忆、工具调用的能力，通过这 3 个能力，使得大模型进化为可以独立实现具体业务逻辑和流程的智能体。AI 智能体技术主要侧重于交互性和用于体验，完成具体流程的实现。

3　算法应用中的挑战、展望

如前所述，基于深度学习的计算机视觉模型已广泛应用于电力场景，为多种多样的作业行为分析任务提供了技术支撑。目标检测、人体姿态估计、图像分类、目标跟踪、视频理解与时空预测等多种算法共同组合成复杂的智能分析系统，从静态图像和动态视频两个维度对监控内容进行理解与分析，极大促进了电力数字化、智能化的发展进程。

3.1　应用难点/挑战

虽然计算机视觉技术在电力作业人员行为分析中进行了大量研究，但在实际应用场景中依然存在诸多难点与挑战。

1）运行的高效性

在大数据、大模型时代，视觉模型对算力的需求越来越高，而电力监控场景的计算资源有限，甚至需要在终端设备进行推理。为此，需要考虑算法

的运行效率从而对模型方案进行针对性设计。与此同时，一些提升模型运行效率的方法，如知识蒸馏、模型剪枝、参数量化等，也值得进一步应用和探索。

2）泛化的鲁棒性

实际应用场景的多样性对算法提出更加鲁棒的要求。开发算法往往需要部署在不同的场景，但不同的变电站环境存在区别，甚至同一变电站的不同摄像头也有拍摄角度、清晰度的差异。此外，不同天气、季节也对算法的鲁棒性提出更高要求。由此，模型的泛化能力也是视觉算法规模化应用的关键。

3）交互的灵活性

电力作业人员行为分析的相关规章、规程需要人为定义，例如“禁行区域”具体位置定义，“聚众拥挤”的人数定义，以及规章制度的迭代和修改。复杂违规行为的定义也需要多种基元行为定义的相互组合。为满足多种多样且富含变化的用户需求，系统交互的灵活性也是值得讨论和探索的课题。

3.2 未来展望

下述方法为应对以上难点和挑战提供了思路：

1）软硬件结合。为解决算法运行高效性的问题，软硬件结合也是具有前景的研究方向。一方面可以研究特定的边缘设备对算法进行加速，以及从算法层面针对特定设备进行结构设计。另一方面，可以将一部分视觉感知功能通过硬件设备解决。

2）通用大模型。提升算法泛化能力最有效的方法是提升训练数据的规模。针对特定场景、特定任务的训练数据有限问题，结合多任务学习方法可提升模型通用视觉表达能力。另外，如何有效利用巨量视频监控无标注数据也是未来研究方向之一。

3）生成式智能。以大语言模型和视觉扩散模型为基础的生成式人工智能，已成为当前最受关注的前沿方向。训练电力领域的模态大语言模型将对算法的交互模式进行革命性提升，也是未来技术发展的重要方向。

4 结论

1）本文深入探讨了基于计算机视觉的电力作业人员行为分析的研究现状与发展趋势。通过细致分析电力作业行为的复杂性，对静态行为分析和动态行为分析的算法模块进行了全面综述，包括目标检测、姿态估计、视频跟踪等核心算法，并针对算法在电力作业行为识别中的高效性、鲁棒性、灵活性等方面的应用难点和挑战进行了深入讨论。

2）随着计算机视觉技术的不断进步和大模型时代的到来，电力视觉领域将迎来新的发展机遇。电力视觉大模型有望实现电力全面监测和智能化维护。同时，电力视觉领域的研究者需要思考如何利用大模型突破现有技术瓶颈，以及如何结合电力行业的实际需求，开发出更加高效、鲁棒、灵活的智能监控系统。

参考文献 References

[1] 戴 彦，王刘旺，李 媛，等. 新一代人工智能在智能电网中的应用研究综述[J]. 电力建设，2018，39(10)：1-11.
DAI Yan, WANG Liuwang, LI Yuan, et al. A brief survey on applications of new generation artificial intelligence in smart grids[J]. Electric Power Construction, 2018, 39(10): 1-11.

[2] 林其雄，陈 畅，闫云凤，等. 一种基于特征引导的电力施工场景工装合规穿戴二阶段检测算法[J]. 浙江电力，2022，41(4)：44-50.
LIN Qixiong, CHEN Chang, YAN Yunfeng, et al. A two-stage detection algorithm for workwear compliance in power construction scenarios based on feature guidance[J]. Zhejiang Electric Power, 2022, 41(4): 44-50.

[3] 彭明智，许 尧，胡永波，等. 基于人工智能技术的变电站二次设备智能巡检技术[J]. 高电压技术，2023，49(增刊 1)：90-96.
PENG Mingzhi, XU Yao, HU Yongbo, et al. Intelligent inspection technology of substation secondary equipment based on artificial intelligence[J]. High Voltage Technology, 2023, 49(Supplement 1): 90-96.

[4] 王刘旺，周自强，林 龙，等. 人工智能在变电站运维管理中的应用综述[J]. 高电压技术，2020，46(1)：1-13.
WANG Liuwang, ZHOU Ziqiang, LIN Long, et al. Review on artificial intelligence in substation operation and maintenance management[J]. High Voltage Engineering, 2020, 46(1): 1-13.

[5] 王 波，张迎晨，齐冬莲，等. 数字化安全管控视角下的全息影像：定义、基本框架及关键技术[J]. 高电压技术，2023，49(8)：3335-3345.
WANG Bo, ZHANG Yingchen, QI Donglian, et al. Holographic images for digital security control: definition, basic framework and key technologies[J]. High Voltage Engineering, 2023, 49(8): 3335-3345.

[6] 任宇路，刘毅敏，节连彬，等. 基于智能视觉的变电站内人员异常行为辅助监控系统[J]. 自动化技术与应用，2024，43(1)：66-70，107.
REN Yulu, LIU Yimin, JIE Lianbin, et al. Auxiliary monitoring system for abnormal behavior of substation personnel based on intelligent video[J]. Techniques of Automation & Applications, 2024, 43(1): 66-70, 107.

[7] 朱惠玲. 基于卷积神经网络的变电站安全监察图像目标识别及定位[D]. 广州：华南理工大学，2021.
ZHU Huiling. Identification and location of substation safety monitoring image based on convolutional neural network[D]. Guangzhou, China: South China University of Technology, 2021.

[8] 张 君. 基于 3D 卷积网络与姿态估计的牵引变电所人体行为识别算法研究[D]. 成都：西南交通大学，2022.
ZHANG Jun. Research on Human behavior recognition algorithm of traction substation based on 3D convolutional network and pose estimation[D]. Chengdu, China: Southwest Jiaotong University, 2022.

[9] 何国立. 基于视频图像的变电站安全违规行为识别算法研究与应用[D]. 杭州：浙江大学，2021.

HE Guoli. Research and application of recognition algorithms for safety violation behaviors in substation based on video images[D]. Hangzhou, China: Zhejiang University, 2021.

[10] HE G L, QI D L. A keypoint-guided pipeline for safety violation identification[C] // 2020 39th Chinese Control Conference (CCC). Shenyang, China: IEEE, 2020: 7223-7228.

[11] HAN Y F, HAN R, LIU L, et al. A real-time video detection method of the surge arrester handling mode based on Yolo-v3 and Open-Pose[C] // 2019 Chinese Control Conference (CCC). Guangzhou, China: IEEE, 2019: 8556-8561.

[12] 何国立，齐冬莲，闫云凤. 一种基于关键点检测和注意力机制的违规着装识别算法及其应用[J]. 中国电机工程学报，2022，42(5)：1826-1836.
HE Guoli, QI Donglian, YAN Yunfeng. An illegal dress recognition algorithm based on key-point detection and attention mechanism and its application[J]. Proceedings of the CSEE, 2022, 42(5): 1826-1836.

[13] 姚宗强，陈宏利，闫文虎，等. 基于视频图像的电力基建施工违章行为自动识别系统设计[J]. 自动化与仪器仪表，2021(3)：143-145，149.
YAO Zongqiang, CHEN Hongli, YAN Wenhu, et al. Design of automatic identification system based on video images for electrical infrastructure construction violations[J]. Automation & Instrumentation, 2021(3): 143-145, 149.

[14] 陈鑫强. 基于深度学习的作业人员违章行为检测技术研究[D]. 厦门：厦门理工学院，2022.
CHEN Xinqiang. Research on detection technology of workers' violation behavior based on deep learning[D]. Xiamen, China: Xiamen University of Technology, 2022.

[15] CI Y Z, WANG Y Z, CHEN M L, et al. UniHCP: A unified model for human-centric perceptions[C] // Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 17840-17852.

[16] 陈　汐，韩译锋，闫云凤，等. 目标物智能跟踪与分割融合算法及其在变电站视频监控中的应用[J]. 中国电机工程学报，2020，40(23)：7578-7586.
CHEN Xi, HAN Yifeng, YAN Yunfeng, et al. A unified algorithm for object tracking and segmentation and its application on intelligent video surveillance for transformer substation[J]. Proceedings of the CSEE, 2020, 40(23): 7578-7586.

[17] 张　楠，李温静，刘　彩，等. 基于多源数据的电力作业人员实时行为安全预警[J]. 计算机与现代化，2023(10)：84-91.
ZHANG Nan, LI Wenjing, LIU Cai, et al. Real-time behavioral safety warning for power operators based on multi-source data[J]. Computer and Modernization, 2023(10): 84-91.

[18] CHEN X, LI Z X, YUAN Y, et al. State-aware tracker for real-time video object segmentation[C] // Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 9381-9390.

[19] 张天予，闵巍庆，韩鑫阳，等. 视频中的未来动作预测研究综述[J]. 计算机学报，2023，46(6)：1315-1338.
ZHANG Tianyu, MIN Weiqing, HAN Xinyang, et al. A survey on future action anticipation in videos[J]. Chinese Journal of Computers, 2023, 46(6): 1315-1338.

[20] NIE X S, CHEN X, JIN H Y, et al. Triplet attention transformer for spatiotemporal predictive learning[C] // Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2024: 7021-7030.

[21] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C] // Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2015: 91-99.

[22] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 779-788.

[23] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]arXiv preprint arXiv: 2010.11929, 2020.

[24] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C] // 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 213-229.

[25] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C] // Proceedings of the 38th International Conference on Machine Learning. Hawaii, USA: PMLR, 2021: 8748-8763.

[26] GIRSHICK R. Fast R-CNN[C] // Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 1440-1448.

[27] 魏业文，李　梅，解园琳，等. 基于改进 Faster-RCNN 的输电线路巡检图像检测[J]. 电力工程技术，2022，41(2)：171-178.
WEI Yewen, LI Mei, XIE Yuanlin, et al. Transmission line inspection image detection based on improved Faster-RCNN[J]. Electric Power Engineering Technology, 2022, 41(2): 171-178.

[28] 赵振兵，熊　静，徐厚东，等. 融合结构推理的深度模型输电线路金具及其缺陷检测[J]. 高电压技术，2023，49(8)：3346-3353.
ZHAO Zhenbing, XIONG Jing, XU Houdong, et al. Integrating structural reasoning for deep model transmission line fittings and their defects detection[J]. High Voltage Engineering, 2023, 49(8): 3346-3353.

[29] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C] // Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2980-2988.

[30] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: towards balanced learning for object detection[C] // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 821-830.

[31] 齐冬莲，韩译锋，周自强，等. 基于视频图像的输变电设备外部缺陷检测技术及其应用现状[J]. 电子与信息学报，2022，44(11)：3709-3720.
QI Donglian, HAN Yifeng, ZHOU Ziqiang, et al. Review of defect detection technology of power equipment based on video images[J]. Journal of Electronics & Information Technology, 2022, 44(11): 3709-3720.

[32] HAN Y F, CHEN X, ZHANG S J, et al. iNL: implicit non-local network[J]. Neurocomputing, 2022, 482: 50-59.

[33] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C] // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 6517-6525.

[34] 刘　江，关向雨，温跃泉，等. 基于改进 YOLOv4 的 GIS 红外特征识别与温度提取方法[J]. 电力工程技术，2023，42(1)：162-168.
LIU Jiang, GUAN Xiangyu, WEN Yuequan, et al. Infrared feature recognition and temperature extraction method of GIS components based on improved YOLOv4[J]. Electric Power Engineering Technology, 2023, 42(1): 162-168.

[35] 李利荣，张云良，陈　鹏，等. 基于上下文信息增强与特征细化的绝缘子破损检测方法[J]. 高电压技术，2023，49(8)：3405-3414.
LI Lirong, ZHANG Yunliang, CHEN Peng, et al. Detection method of

insulator breakage based on context augmentation and feature refinement[J]. High Voltage Engineering, 2023, 49(8): 3405-3414.

[36] WANG Y Z, CHEN M L, TANG S X, et al. Unsupervised object detection pretraining with joint object priors generation and detector learning[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022: 905.

[37] TIAN Z, SHEN C H, CHEN H, et al. FCOS: Fully convolutional one-stage object detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 9626-9635.

[38] DUAN K W, BAI S, XIE L X, et al. CenterNet: Keypoint triplets for object detection[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 6568-6577.

[39] 别一凡，李 波，江 军，等. 基于改进 SSD 的变压器套管红外图像油位智能识别方法[J]. 电力工程技术，2021，40(5)：158-163.
BIE Yifan, LI Bo, JIANG Jun, et al. Intelligent oil level recognition of transformer bushing infrared image based on improved SSD algorithm[J]. Electric Power Engineering Technology, 2021, 40(5): 158-163.

[40] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multiBox detector[C]// 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 21-37.

[41] FAN Z Z, SHI L R, XI C, et al. Real time power equipment meter recognition based on deep learning[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 5017015.

[42] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 9992-10002.

[43] 李帷韬，侯建平，张 倩，等. 基于强化学习和 Transformer 的输电线路缺陷智能检测方法研究[J]. 高电压技术，2023，49(8)：3373-3384.
LI Weitao, HOU Jianping, ZHANG Qian, et al. Research on intelligent detection method of transmission line defects based on reinforcement learning and transformer[J]. High Voltage Engineering, 2023, 49(8): 3373-3384.

[44] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv: 2010.04159, 2020.

[45] 宰红斌，吴浩林，王 昊，等. 基于改进机器学习的输电线路弧垂温度估计方法[J]. 电力工程技术，2022，41(2)：209-214，223.
ZAI Hongbin, WU Haolin, WANG Hao, et al. Sag and temperature estimation method based on improved machine learning for transmission line[J]. Electric Power Engineering Technology, 2022, 41(2): 209-214, 223.

[46] LIU X L, WANG Q M, HU Y, et al. End-to-end temporal action detection with transformer[J]. IEEE Transactions on Image Processing, 2022, 31: 5427-5441.

[47] DOSOVITSKIY A, KOLTUN V. Learning to act by predicting the future[C]// 5th International Conference on Learning Representations. Toulon, France: OpenReview. net, 2017.

[48] LI L H, ZHANG P C, ZHANG H T, et al. Grounded language-image pre-training[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 10955-10965.

[49] LIU S L, ZENG Z Y, REN T H, et al. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection[J]. arXiv preprint arXiv: 2303.05499, 2023.

[50] ZHOU X Y, GIRDHAR R, JOULIN A, et al. Detecting twenty-thousand classes using image-level supervision[C]// 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 350-368.

[51] ZANG Y H, LI W, ZHOU K Y, et al. Open-vocabulary DETR with conditional matching[C]// 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 106-122.

[52] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Proceedings of the 2016 IEEE conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016: 4724-4732.

[53] NEWELL A, YANG K Y, DENG J. Stacked hourglass networks for human pose estimation[C]// 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 483-499.

[54] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 5686-5696.

[55] XU Y F, ZHANG J, ZHANG Q M, et al. ViTPose: simple vision transformer baselines for human pose estimation[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022: 2795.

[56] FANG H S, XIE S Q, TAI Y W, et al. RMPE: regional multi-person pose estimation[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 2353-2362.

[57] PAPANDREOU G, ZHU T, CHEN L C, et al. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model[C]// Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 2018: 282-299.

[58] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017: 1302-1310.

[59] KREISS S, BERTONI L, ALAHI A. PifPaf: composite fields for human pose estimation[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 11969-11978.

[60] JIN S, LIU W T, XIE E Z, et al. Differentiable hierarchical graph grouping for multi-person pose estimation[C]// 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 718-734.

[61] YANG J, ZENG A L, LIU S L, et al. Explicit box detection unifies end-to-end multi-person pose estimation[C]// The Eleventh International Conference on Learning Representations. Kigali, Rwanda: OpenReview. net, 2023.

[62] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A skinned multi-person linear model[J]. Seminal Graphics Papers: Pushing the Boundaries, 2023, 2: 88.

[63] GÜLER R A, NEVEROVA N, KOKKINOS I. DensePose: dense human pose estimation in the wild[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7297-7306.

[64] PAVLAKOS G, CHOUTAS V, GHORBANI N, et al. Expressive body capture: 3D hands, face, and body from a single image[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 10967-10977.

[65] SANTESTEBAN I, GARCES E, OTADUY M A, et al. SoftSMPL: data-driven modeling of nonlinear soft-tissue dynamics for parametric humans[J]. Computer Graphics Forum, 2020, 39(2): 65-75.

[66] OSMAN A A A, BOLKART T, BLACK M J. STAR: sparse trained articulated human body regressor[C]// 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 598-613.

[67] WANG H Y, GÜLER R A, KOKKINOS I, et al. BLSM: a bone-level skinned model of the human mesh[C]// 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 1-17.

[68] XU H Y, BAZAVAN E G, ZANFIR A, et al. GHUM & GHUML: Generative 3D human shape and articulated pose models[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020: 6183-6192.

[69] BEWLEY A, GE Z Y, OTT L, et al. Simple online and realtime tracking[C]// 2016 IEEE International Conference on Image Processing (ICIP). Phoenix, USA: IEEE, 2016: 3464-3468.

[70] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]// 2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE, 2017: 3645-3649.

[71] ZHANG Y F, SUN P Z, JIANG Y, et al. ByteTrack: multi-object tracking by associating every detection box[C]// 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022: 1-21.

[72] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[C]// 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 107-122.

[73] ZHANG Y F, WANG C Y, WANG X G, et al. Fairmot: on the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[74] ZHOU X Y, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]// 16th European Conference on Computer Vision. Glasgow, UK: Springer, 2020: 474-490.

[75] SUN P Z, CAO J K, JIANG Y, et al. TransTrack: multiple object tracking with transformer[Z]. arXiv preprint arXiv: 2012.15460, 2020.

[76] MEINHARDT T, KIRILLOV A, LEAL-TAIXÉ L, et al. TrackFormer: multi-object tracking with transformers[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 8834-8844.

[77] DE PLAEN P F, MARINELLO N, PROESMANS M, et al. Contrastive learning for multi-object tracking with transformers[C]// Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE, 2024: 6853-6863.

[78] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2014: 568-576.

[79] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015: 4489-4497.

[80] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 6450-6459.

[81] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016: 20-36.

[82] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 7794-7803.

[83] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 6201-6210.

[84] BERTASIUS G, WANG H, TORRESANI L. Is space-time attention all you need for video understanding?[J]. arXiv preprint arXiv: 2102.05095, 2021.

[85] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: a video vision transformer[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 6816-6826.

[86] ZHANG Y Y, LI X Y, LIU C H, et al. VidTr: video transformer without convolutions[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 13557-13567.

[87] FAN H Q, XIONG B, MANGALAM K, et al. Multiscale vision transformers[C]// Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021: 6804-6815.

[88] SHI X J, CHEN Z R, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 2015: 802-810.

[89] WANG Y B, GAO Z F, LONG M S, et al. Predrnn++: towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning[C]// Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Sweden: PMLR, 2018: 5110-5119.

[90] WANG Y B, WU H X, ZHANG J J, et al. PredRNN: a recurrent neural network for spatiotemporal predictive learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(2): 2208-2225.

[91] GAO Z Y, TAN C, WU L R, et al. SimVP: simpler yet better video prediction[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022: 3160-3170.

[92] TAN C, GAO Z Y, WU L R, et al. Temporal attention unit: towards efficient spatiotemporal predictive learning[C]// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 18770-18782.

[93] HU X T, HUANG Z W, HUANG A L, et al. A dynamic multi-scale voxel flow network for video prediction[C]// Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE, 2023: 6121-6131.

[94] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: open and efficient foundation language models[Z]. arXiv preprint arXiv: 2302.13971, 2023.

[95] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[J]. arXiv preprint arXiv: 2304.08485, 2023.

[96] ZHU D Y, CHEN J, SHEN X Q, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models[Z]. arXiv preprint arXiv: 2304.10592, 2023.

[97] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: a frontier large vision-language model with versatile abilities[Z]. arXiv preprint arXiv: 2308.12966, 2023.

[98] ZHANG S L, SUN P Z, CHEN S F, et al. GPT4RoI: instruction tuning

large language model on region-of-interest[J]. arXiv preprint arXiv: 2307.03601, 2023.

[99] PENG Z L, WANG W H, DONG L, et al. Kosmos-2: grounding multimodal large language models to the world[Z]. arXiv preprint arXiv: 2306.14824, 2023.

[100] CHEN K Q, ZHANG Z, ZENG W L, et al. Shikra: unleashing multimodal LLM's referential dialogue magic[Z]. arXiv preprint arXiv: 2306.15195, 2023.

[101] MAAZ M, RASHEED H, KHAN S, et al. Video-chatGPT: towards detailed video understanding via large vision and language models[J]. arXiv preprint arXiv: 2306.05424, 2023.

[102] LI Y W, WANG C Y, JIA J Y. LLaMA-VID: an image is worth 2 tokens in large language models[Z]. arXiv preprint arXiv: 2311.17043, 2023.

[103] 赵振兵，冯 烁，席 悦，等. 大模型时代：电力视觉技术新起点[J/OL]. 高电压技术，2024：1-14[2024-03-14]. http://kns.cnki.net/kcms/detail/42.1239. tm.20240227.1412.001.html.
ZHAO Zhenbing, FENG Shuo, XI Yue, et al. The era of large models: a new starting point for power vision technology[J/OL]. High Voltage Engineering, 2024: 1-14[2024-03-14]. http://kns.cnki.net/kcms/detail/42.1239. tm.20240227.1412.001.html.



闫云凤
1988—，女，博士，副教授，硕导
主要从事计算机视觉、运行态势感知、分布式发电设备异常监测方面的研究工作
E-mail: 21210004@zjuedu.cn

YAN Yunfeng
Ph.D.
Associate professor



陈 汐
1995—，男，博士生
主要从事生成式人工智能和多模态学习方面的研究工作
E-mail: chauncey0620@gmail.com

CHEN Xi
Ph.D. candidate



金浩远
1997—，男，博士生
主要从事多目标跟踪、视频预测和行为识别方面的研究工作
E-mail: jhyjhy@zju.edu.cn

JIN Haoyuan
Ph.D. candidate



齐冬莲(通信作者)
1973—，女，博士，教授，博导
主要从事智能信息处理、混沌系统、非线性理论与应用方面的研究工作
E-mail: qidl@zju.edu.cn

QI Donglian
Ph.D., Professor
Corresponding author

收稿日期 2024-03-19 修回日期 2024-04-10 编辑 卫李静