IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 February 2024, accepted 4 May 2024, date of publication 9 May 2024, date of current version 30 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3399222

SURVEY

Efficient Monocular Human Pose Estimation Based on Deep Learning Methods: A Survey

XUKE YAN[1], (Member, IEEE), BO LIU[2], (Senior Member, IEEE), AND GUANGZHI QU[1], (Senior Member, IEEE)
[1]Department of Computer Science and Engineering, Oakland University, Rochester, MI 48309, USA
[2]School of Mathematical and Computational Sciences, Massey University, Palmerston North 4472, New Zealand

Corresponding author: Guangzhi Qu (gqu@oakland.edu)

ABSTRACT Human pose estimation (HPE) is a crucial computer vision task with a wide range of applications in sports medicine, healthcare, virtual reality, and human-computer interaction. The demand for real-time HPE solutions necessitates the development of efficient deep-learning models that can be deployed on resource-constrained devices. While a few surveys exist in this area, none delve deeply into the critical intersection of efficiency and performance. This survey reviews the state-of-the-art efficient deep learning approaches for real-time HPE, focusing on strategies for improving efficiency without compromising accuracy. We discuss popular backbone networks for HPE, model compression techniques, network pruning and quantization, knowledge distillation, and neural architecture search methods. Furthermore, we critically analyze the existing works, highlighting their strengths, weaknesses, and applicability to different scenarios. We also present an overview of the evaluation datasets, metrics, and design for efficient HPE. Finally, we identify research gaps and challenges in the field, providing insights and recommendations for future research directions in developing efficient and scalable HPE solutions.

INDEX TERMS Survey, 2D human pose estimation, 3D human pose estimation, deep learning, efficiency.

I. INTRODUCTION

A. MOTIVATION OF HUMAN POSE ESTIMATION

Human Pose Estimation (HPE) has emerged as a fundamental and challenging task in the computer vision community. The primary objective of HPE is to predict human pose information, such as the spatial locations of body joints and/or body shape parameters, from monocular images or videos [1]. HPE holds great significance due to its ability to provide detailed pose information without the need for complex multi-camera setups or wearable markers, making it a crucial component of numerous computer vision tasks [2].

The motivation behind HPE research lies in its potential to bridge the gap between the digital and physical worlds, enabling a deeper understanding of human behavior, movement, and interaction. The rapid development of HPE methods, driven by advances in deep learning technologies and the availability of large-scale 2D/3D pose datasets, has led to significant performance improvements in both accuracy and efficiency. This progress has driven research efforts toward exploring innovative network designs, multi-task interactions, and body model explorations, further enhancing HPE's capabilities [3].

Monocular HPE tasks can be divided into two main categories based on the spatial dimension of the output results: 2D pose estimation and 3D pose estimation. While 2D pose estimation focuses on locating the 2D coordinates of human anatomical keypoints (body joints) in images, 3D pose estimation aims to predict the depth information for a more accurate spatial representation. The intrinsic connections between 2D and 3D pose estimation and the growing demand for detailed pose information have prompted researchers to investigate HPE methods that bridge the gap between 2D and 3D representations [4].

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

Despite the successes achieved in HPE performance and practice, there is still room for improvement and further exploration. Comprehensive reviews of representative algorithms and insightful analyses of 2D-to-3D pose estimation remain limited, highlighting the need for continued research in this domain. As HPE continues to advance, its significance and the motivation behind its development will only grow, paving the way for new and improved applications in various fields.

B. REAL-TIME HPE APPLICATIONS AND THE NEED FOR EFFICIENCY

Real-time human pose estimation (HPE) applications require efficient models that can process data quickly and accurately while being deployed on resource-constrained devices. As shown in Figure 1 The following subsections discuss various real-time HPE applications and the importance of efficiency in these use cases.

FIGURE 1. Real-world examples of various real-time applications: images are from OpenPose HPE demonstration [5].

1) REMOTE PATIENT MONITORING

Remote patient monitoring systems allow healthcare professionals to track patients' vital signs, movements, and activities outside clinical settings [6]. Efficient HPE models can be used to monitor patients' posture, exercise adherence, and physical rehabilitation progress in real-time, providing timely feedback and interventions. Such systems are particularly important during pandemics like COVID-19, where remote monitoring can help minimize in-person interactions and reduce the risk of infection [7].

2) ELDERLY CARE AND FALL DETECTION

Efficient real-time HPE can play a crucial role in elderly care by detecting falls and monitoring daily activities to ensure the well-being of older adults [8]. By identifying abnormal movements, posture changes, or signs of distress, caregivers can intervene quickly, potentially preventing injuries or other adverse events.

3) FITNESS TRACKING AND PERSONAL TRAINING

Fitness trackers and personal training apps often use HPE models to provide real-time feedback on users' exercise form, intensity, and progress. Efficient HPE models can help users maintain proper form during workouts, prevent injuries, and tailor their training programs based on individual needs and goals [9], [10].

4) SURVEILLANCE AND CROWD ANALYSIS

Real-time HPE can be employed in surveillance and crowd analysis applications to detect unusual or potentially dangerous activities. Efficient models can analyze multiple people simultaneously, providing valuable insights into crowd behavior and enabling rapid response to security threats or emergencies [11], [12].

5) PEDESTRIAN DETECTION FOR AUTONOMOUS DRIVING

Among many objects that autonomous vehicles must accurately detect and interpret, pedestrians are among the most crucial. A fast response HPE enhances pedestrian detection systems within these vehicles. By analyzing the pose and posture of individuals, the HPE algorithm [13] allows the system to predict potential pedestrian movements more accurately, whether it's someone about to cross the road or a person momentarily pausing on the sidewalk. Given the need for swift data processing in autonomous driving, lightweight, streamlined, effective models are essential. Such models ensure the vehicle responds promptly to dynamic street scenarios, significantly bolstering road safety.

The ever-evolving landscape of HPE applications, from augmented reality experiences and gaming interfaces to advanced healthcare monitoring and industrial automation, demands real-time processing for seamless user interaction and safety. This real-time requirement implies that any delays, even if minimal, can hinder user experience or even result in potential hazards, especially in applications like patient monitoring or machinery control.

Furthermore, many real-time HPE applications operate in environments where high computational power is not readily available. Consider, for instance, a fitness app on a smartphone providing real-time feedback on an individual's exercise form. The app cannot afford to drain the battery rapidly by demanding extensive computational resources, nor can it compromise on accuracy, which can lead to incorrect feedback. Similarly, edge devices in remote locations, like surveillance cameras in wildlife settings, need to be efficient both in terms of computational demands and energy consumption.

Hence, balancing computational efficiency with model accuracy becomes not just an aspiration but an essential factor in designing HPE systems. It's a challenging intersection where the academic and industrial research communities are investing significant efforts. The call for efficient models aligns not only with the technical constraints but also the aspiration for broader applicability and accessibility of real-time HPE solutions across diverse platforms and scenarios.

C. REVIEW OF PREVIOUS SURVEY WORK ON HPE

In this section, we present an overview of prior survey until 2023 works related to Human Pose Estimation (HPE). Our primary goal is to discuss the contributions of these works while also identifying the research gaps that our current survey is to bridge. Subsequently, Figure 2 offers insights into these previous HPE surveys, their focus, and the gaps they leave.

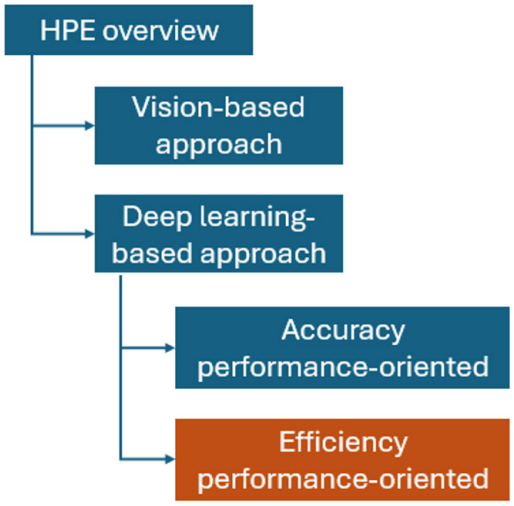

FIGURE 2. A diagram of previous HPE surveys and ours.

The previous surveys [14], [15] have made significant contributions to summarizing and reviewing the advancements in HPE. However, those employing deep learning methods [3], [4], [16], [17] have mainly focused on either 2D and 3D HPE methods. And they mainly emphasize on improving the detection accuracy. These surveys have not addressed the efficiency of HPE methods, which is crucial when considering real-time applications and environments constrained by computational resources.

Our survey fills this gap by providing a comprehensive review of efficient HPE methods based on deep learning, covering both 2D and 3D HPE approaches. By focusing on the efficiency aspect of HPE, we aim to offer valuable insights to researchers and practitioners working on real-time HPE applications, such as mobile, edge computing, and remote monitoring.

D. CONTRIBUTIONS

This survey aims to provide a comprehensive and updated overview of efficient Human Pose Estimation (HPE) methods based on deep learning, with a focus on both 2D and 3D HPE approaches. Our contributions are as follows:

- We consider both efficient 2D and 3D HPE methods: Unlike previous surveys that focused on either 2D or 3D HPE separately, we provide a holistic view of the efficient HPE methods in both domains. This allows researchers and practitioners to better understand the connections between 2D and 3D HPE and explore potential synergies in their work.
- We emphasize the efficiency of HPE methods: While previous surveys have mainly concentrated on improving detection accuracy, we specifically focus on the efficiency aspect of HPE methods. This is particularly important for real-time applications and resource-constrained environments, such as mobile devices, edge computing, and remote monitoring.
- We analyze and explore various approaches to improve efficiency: We present a detailed analysis of various techniques employed to enhance the efficiency of HPE methods, such as model compression techniques, network pruning and quantization, knowledge distillation, architecture search (NAS), and smaller baseline network designs with improved accuracy.
- We review of evaluation datasets and metrics: We provide an extensive overview of the datasets and evaluation metrics used in HPE research, which is crucial for understanding the progress and challenges in the field.
- We identify future research directions: We discuss the challenges and limitations of existing efficient HPE methods and propose promising future research directions. This analysis can inspire and guide researchers working on HPE to explore new ideas and solutions.

By offering a comprehensive, up-to-date (until 2024), and focused review of efficient HPE methods, we believe that this survey will serve as a valuable resource for researchers, practitioners, and students interested in HPE and its real-world applications.

E. PAPER ORGANIZATION

The structure of this survey paper unfolds as follows: After this introduction, Section II explores deep learning frameworks employed in both 2D and 3D HPE, along with their associated applications. Section III delves into methodologies for efficient deep learning. Section IV outlines evaluation metrics pertinent to HPE efficiency and introduces commonly used 2D/3D datasets. In Section V, we holistically compare various methods within 2D and 3D efficient HPE. Section VI presents prevailing challenges, unresolved queries, and potential avenues for future research.

II. DEEP LEARNING FRAMEWORKS FOR HPE

Adapting these foundational deep learning architectures for both 2D and 3D HPE showcases the versatility and depth of these models. The nuances of each method highlight the increasing specificity and accuracy of pose estimation techniques in modern computer vision applications.

A. CONVOLUTIONAL NEURAL NETWORKS (CNNs)

CNNs [18] harness specialized layers, namely convolutional layers, that are adept at learning spatial hierarchies from image data.

1) APPLICATIONS IN 2D HPE

CNNs process images to detect 2D coordinates of human anatomical key points in images. Fine-tuned CNNs like VGG and ResNet process image patches around body joints to determine their 2D positions in images. Wei et al. [19] present Convolutional Pose Machines (CPMs), a method that integrates convolutional networks into a sequential prediction framework to improve 2D human pose estimation. This approach learns image features and spatial models directly from data and demonstrates state-of-the-art performance on benchmarks such as MPII, LSP, and FLIC, marking a significant advancement in the early development of deep learning-based HPE. Singh et al. [20] review 2D human pose estimation challenges, evaluates various research methods, introduces a CNN-based model, and suggests potential future research directions. Toshev and Szegedy [21] introduce a holistic human pose estimation method utilizing Deep Neural Networks, which offers precise joint localization through a cascading DNN approach, leveraging the advantages of deep learning for this task.

2) APPLICATIONS IN 3D HPE

For 3D HPE, CNNs often use depth maps or employ a regression mechanism to infer depth for each joint, converting 2D joint positions into 3D coordinates. Mehta et al. [22] introduce a CNN-based technique for 3D human pose estimation from single RGB images, leveraging both existing 2D and 3D pose data for improved generalizability. They also present a new diverse dataset for pose estimation and highlight the significance of transfer learning for achieving robust results in real-world scenarios [23]. A novel method is proposed by Ghezelghieh, et al. to estimate 3D human pose from a single RGB image by leveraging camera viewpoint in tandem with 2D joint locations, utilizing a trained CNN and 3D computer rendering, resulting in a significant error reduction on the Human3.6m benchmark.

B. STACKED HOURGLASS NETWORKS

This innovative network captures multi-scale information using a repeated bottom-up, top-down processing structure.

1) APPLICATIONS IN 2D HPE

Designed to capture information at various scales, this network works effectively for 2D HPE by focusing on spatial hierarchies and estimating joint locations in images at multiple resolutions. Newell et al. [24] firstly presents a "stacked hourglass" convolutional network architecture for human pose estimation that utilizes repeated bottom-up, top-down processing with intermediate supervision. The network is based on the successive steps of pooling and upsampling that are done to produce a final set of predictions, which achieves state-of-the-art results.

2) APPLICATIONS IN 3D HPE

Extensions of the stacked hourglass model have been used to predict 3D skeletal joint positions. The multi-scale processing assists in extracting depth information from 2D images. Xu and Takano [25] present the Graph Stacked Hourglass Networks, a unique graph convolutional design for 2D-to-3D human pose estimation, emphasizing multi-scale skeletal representations and deep multi-level features to enhance estimation accuracy.

C. HIGH-RESOLUTION NETWORK (HRNet)

Unlike traditional models that downsample and then upsample, HRNet maintains high-resolution representations through parallel multi-resolution convolutions.

1) APPLICATIONS IN 2D HPE

By maintaining high-resolution representations throughout, HRNet captures minute details essential for accurate 2D pose estimation, especially in high-resolution images. Sun et al. [26] introduce a unique approach to human pose estimation that consistently maintains high-resolution representations, initiates with a high-resolution subnetwork, and progressively integrates high-to-low resolution subnetworks, utilizing multi-scale fusions for enriched representation throughout the process.

2) APPLICATIONS IN 3D HPE

There is no known application of HRNet in 3D HPE to our knowledge

D. TRANSFORMERS

Transformers have recently gained prominence in both 2D and 3D Human Pose Estimation (HPE) due to their ability to capture long-range dependencies and global evidence of keypoints, outperforming conventional CNNs in certain aspects.

1) APPLICATIONS IN 2D HPE

The advent of transformer-based models for 2D HPE has introduced new capabilities in capturing fine-grained evidence and overcoming occlusions. An early example, TransPose [27], leveraged attention layers to predict keypoint heatmaps, effectively handling occlusion scenarios. Subsequent models like TokenPose [28] utilized token representations to capture constraint cues and visual relationships, enhancing the understanding of keypoint configurations. HRFormer [29] introduced a high-resolution approach by integrating transformer modules into HRNet, improving memory and computational efficiency. The Token-Pruned Pose Transformer (PPT) [30] further advanced the field by efficiently estimating poses and enabling direct instance-aware body pose estimation.

2) APPLICATIONS IN 3D HPE

In the realm of 3D HPE, transformer integration has been explored to regress SMPL mesh vertices from single images, combining CNNs with transformers for improved accuracy but at the cost of higher computational and memory demands, as seen in METRO [31] and MeshGraphormer [32]. Efforts to mitigate these costs have led to the development of lightweight transformer architectures like FeatER [33], which significantly reduce the parameter count and computational overhead while outperforming METRO in efficiency, marking a critical step towards scalable and effective 3D human pose estimation.

III. IMPROVING EFFICIENCY FOR HPE

Improving the efficiency of Human Pose Estimation (HPE) is essential, especially in applications necessitating real-time response or deployment in resource-constrained environments. Here's a deep dive into various strategies to optimize HPE models for efficiency:

A. NETWORK PRUNING AND QUANTIZATION

Pruning eliminates less important neurons or connections, significantly reducing model size and computational demand. In conjunction, the Quantization of neural networks involves converting a model's continuous-valued weights and activations to a discrete set of values, often represented with lower precision. This is achieved by mapping the full-precision values to a set of fixed levels, typically within a lower bit-width format, thereby compressing the model. This reduces the memory footprint and computational requirements, allowing for faster model operations and deployment in resource-limited settings.

For HPE, the synergy of reduced computational complexity through pruning and quantization means faster inference times, a cornerstone for real-time pose estimation. Bulat and Tzimiropoulos [34] utilize the binarized network based on an Hourglass network, and convert most of the parameters into -1 or 1. The proposed hierarchical, parallel, and multi-scale residual architecture can improve the accuracy over their bottleneck blocks but significantly reduce the model's size, which makes it accommodate resource-limited platforms.

B. KNOWLEDGE DISTILLATION

Knowledge Distillation [35], [36] is a model compression technique where a smaller model, often referred to as the student, is trained to mimic the behavior of a larger model, known as the teacher. Instead of learning from the original ground truth labels, the student model is trained on the softer output distributions (probabilities) of the teacher model. This softer distribution, which might contain information about the relationships between different classes, can be more informative than the hard labels, allowing the student to achieve better performance than if it were trained directly on the ground truth. A knowledge distillation system is composed of three key components: knowledge, distillation algorithm, and teacher-student architecture, as shown in Figure 3.

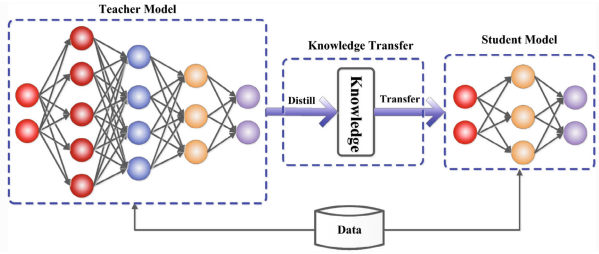

FIGURE 3. A general teacher-student framework for knowledge distillation [36].

In the context of Human Pose Estimation (HPE), by distilling the knowledge from a large, accurate, but computationally intensive HPE model (teacher) into a smaller, faster model (student), it's possible to retain much of the performance of the larger model while benefiting from the increased efficiency of the smaller one as presented by Zhang et al. [37]. This makes the student model more suitable for real-time pose estimation tasks, especially on edge devices. Furthermore, the distilled model, being lighter, consumes less memory and computational power, making it more energy-efficient and faster in its predictions. This amalgamation of speed and accuracy through Knowledge Distillation can significantly elevate the efficacy of HPE applications.

C. NEURAL ARCHITECTURE SEARCH (NAS)

NAS [38] is a method that automates the process of selecting the best neural network architecture for a specific task. To achieve this, NAS employs search strategies, such as reinforcement learning or evolutionary algorithms, to navigate through the predefined architectural space. Over a series of trials, it evaluates the performance of various architectures on the task, refining its search based on the results. As it progresses, the search refines based on performance outcomes, allowing NAS to algorithmically pinpoint the optimal architecture rather than relying on manual design. As shown in Figure 4, A search strategy selects an architecture A from a predefined search space a. The architecture is passed to a performance estimation strategy, which returns the estimated performance of A to the search strategy.

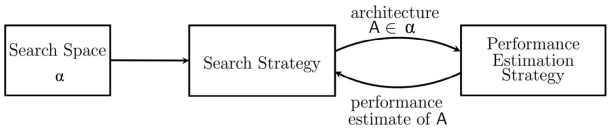

FIGURE 4. A workflow of Neural Architecture Search methods [38].

Within the context of HPE, NAS can lead to the discovery of architectures that inherently balance efficiency and accuracy. Given the constraints set during the search,

custom architectures might emerge that are more adept at handling HPE's unique challenges without the overhead of unnecessary parameters. Tan and Le [39] introduce an EfficientNet achieving unparalleled accuracy and efficiency with substantially fewer parameters. This method scales ConvNets uniformly across depth, width, and resolution, enhancing existing compact model performance. Bao et al. [40] propose a model named PoseNAS that utilizes a NAS-driven method to seek out a data-focused pose network comprising stacked searchable units, optimizing both feature extraction and fusion specifically for pose-related tasks. Xu et al. [41] present a spatial network by methodically structuring a search realm across five varied parameters: network depth, width, kernel dimension, group count, and attention mechanisms. This approach is extended to video pose estimation, pinpointing temporal feature fusion and auto-calculating allocations within videos.

D. MORE COMPACT NETWORK DESIGN

Starting with a minimalist baseline network and then carefully integrating modules to enhance accuracy forms a cornerstone for achieving efficiency without compromising on performance. This tailored approach renders the model adaptable across a broad spectrum of applications, from mobile devices to extensive cloud-based systems, while maintaining competitive accuracy with a reduced computational footprint.

Efficient architectures like MobileNets [42] and SqueezeNets [43] play a pivotal role. MobileNets, for example, utilize depth-wise separable convolutions to significantly cut down the number of parameters and computational complexity, facilitating the deployment of HPE models on devices with limited processing power without sacrificing estimation quality. Incorporating attention mechanisms, which have shown substantial success in natural language processing tasks, into compact HPE models can further refine their ability to capture the nuanced spatial relationships between body joints. The integration of transformer models, known for their proficiency in handling long-range dependencies [44], into smaller network designs could elevate the performance of HPE systems in complex scenarios. Moreover, the adaptation of compact transformer architectures, such as those seen in recent studies for visual tasks, could potentially minimize the computational demands typically associated with transformers while retaining their effectiveness in capturing intricate patterns within data [45].

IV. DATASETS AND EVALUATION METRICS

A. EFFICIENCY EVALUATION METRICS

1) THE NUMBER OF PARAMETERS

The number of parameters in a deep learning model is a metric for its efficiency for various reasons. Firstly, it directly relates to the model's size in storage, making it crucial for deployment in storage-constrained devices like smartphones or embedded devices [34], [46]. Secondly, a model with fewer parameters is computationally efficient in training and inference, making it more time-efficient [47]. Additionally, lightweight models are more energy-efficient, a critical factor for battery-powered devices [48]. However, it's essential to strike a balance. Reducing the number of parameters can make a model more efficient, but if the model becomes too simple, it might not capture the complexities of the data, leading to underfitting. The optimal number of parameters is highly problem-dependent and often discovered through experimentation and iterative refinement.

2) FLOATING POINT OPERATIONS

The FLOP (Floating Point Operations) evaluates the computational cost of deep learning models by quantifying the number of operations required to generate an output. This measure is valuable as it provides insights into a model's inference speed, which is especially vital for real-world applications. FLOP also helps determine a model's compatibility with devices with limited computational resources [49].

B. DATASETS FOR 2D HPE

Starting from 2014, various datasets were used for 2D HPE tasks. However, most recent studies have shifted away from these early datasets due to their limitations, such as the lack of varied object movements and limited data quantity. Since deep learning approaches thrive on vast training data, we focus on contemporary large-scale datasets for 2D human pose estimation.

1) MPII HUMAN POSE DATASET [2]

This dataset, created by the Max Planck Institute for Informatics, stands as a benchmark for evaluating articulated HPE. It comprises approximately 25,000 images showcasing over 40,000 individuals with marked body joints. Annotations in MPII are comprehensive, covering aspects like body part occlusions, 3D torso, and head orientations, all of which were labeled through Amazon Mechanical Turk. This dataset is especially apt for 2D single or multi-person HPE evaluations.

2) MICROSOFT COCO DATASET [50]

Undoubtedly one of the most utilized large-scale datasets, COCO encompasses over 330,000 images with more than 200,000 labeled subjects having keypoints. Each labeled person features 17 joint annotations. Expanding on this, Jin et al. [51] introduced the COCO-WholeBody Dataset, enhancing the annotations to capture the entire human body.

3) PoseTrack DATASET [52]

This dataset is specifically designed for HPE and articulated tracking within video formats, especially addressing the challenges posed by body part occlusion and truncation in densely populated settings. It contains 1,138 video sequences with 153,615 pose annotations, split into 593 for training, 170 for validation, and 375 for testing. Each individual in PoseTrack is annotated with 15 joints, supplemented by a keypoint visibility label.

TABLE 1. Datasets for 2D HPE.

Dataset	Year	Type	Image/Video	Num of joints	Number of images or Videos			Evaluation
					Train set	Val set	Test set	
LSP [53]	2010	Single	Image	14	1k	-	1k	PCK
FLIC [54]	2013	Single	Image	10	5k	-	1k	PCK
Penn Action [55]	2013	Single	Video	13	1k	-	1k	PCK
MPII Single [2]	2014	Single	Image	16	29k	-	12k	PCK
MPII Multiple [2]	2014	Multiple	Image	16	3.8k	-	1.7k	PCK
COCO [50]	2016	Multiple	Image	17	45k	22k	80k	AP
AIC-HKD [56]	2017	Multiple	Image	14	210k	30k	60k	mAP
PoseTrack [52]	2018	Multiple	Video	13	593	170	375	mAP
CrowdPose [57]	2019	Multiple	Image	14	10k	2k	8k	mAP
HiEve [58]	2020	Multiple	Video	14	19	-	13	mAP

More datasets such as LSP [53], FLIC [54], Penn Action [55] for single-person dataset, and AIC-HKD [56], CrowdPose [57], HiEve [58] for multiple people datasets are listed in Table 1, with the year of publication, type, number of joints, number of datasets and the evaluation methods.

C. EVALUATION METRICS FOR 2D HPE

Evaluating 2D HPE is complex due to the diverse range of factors to consider, such as human body size, single vs. multiple pose estimation, and focus on upper or full body parts. This has led to the development of numerous metrics, and we outline some predominant ones below:

1) PERCENTAGE OF CORRECT PARTS (PCP) [59]

Historically prevalent in early 2D HPE literature, PCP gauges the accuracy of limb localizations. A limb's localization is deemed accurate if the distance between the estimated and actual joint is within a specified fraction (usually between 0.1 to 0.5) of the limb's length. In certain contexts, this metric is labeled as PCP@0.5, indicating a 0.5 threshold. Notably, PCP focuses on single-person HPE, but its usage has declined as it unfairly penalizes harder-to-detect shorter limbs.

2) PERCENTAGE OF CORRECT KEYPOINTS (PCK) [60]

This metric evaluates the precision of keypoint localizations against a set threshold, commonly 50% of the head segment's length in a test image, termed PCKh@0.5. Another variant, PCK@0.2, is used when the distance between detected and actual joints is less than 0.2 times the torso's diameter. Higher PCK values indicate superior model efficacy.

3) AVERAGE PRECISION (AP) AND AVERAGE RECALL (AR) [50]

AP is a metric that assesses keypoint detection accuracy based on precision (the fraction of accurate positive results out of all positive predictions) and recall (the fraction of accurate positive predictions out of all actual positive instances). AP calculates the mean precision for recall values ranging from 0 to 1. Some variations of this metric include Average Precision of Keypoints (APK) and Mean Average Precision (mAP), popular on datasets like MPII and PoseTrack. In contrast, the AR metric, prominent in the COCO keypoint [50] evaluation, focuses on recall. Object Keypoint Similarity (OKS), analogous to Intersection over Union (IoU) in object detection contexts, is incorporated for both AP and AR evaluations. COCO evaluations typically utilize mAP across 10 OKS thresholds.

D. DATASETS FOR 3D HPE

Acquiring accurate 3D annotation for 3D HPE datasets is a challenging task, compared to 2D HPE datasets. Datasets for 3D HPE primarily originate from motion capture systems or inertial measurement units (IMU), while some recent datasets are generated from game engines. we'll discuss some prominent datasets, including recent ones, and provide an overview in Table 2.

1) HUMAN3.6M [61]

Standing out as one of the most expansive motion capture collections, it boasts 3.6 million human poses paired with corresponding images. It provides precise 3D human joint positions and synchronized high-resolution videos recorded at 50 Hz. The dataset captures 11 professional actors performing 17 scenarios from four distinct camera angles. Human3.6M dataset is split into two protocols named Protocol 1 and Protocol 2. Protocol 1 consists of S1, S5, S6, S7, S8, S9 for training and S11 for testing. Protocol 2 contains S1, S5, S6, S7, S8 for training and S9, S11 for testing.

2) MPI-INF-3DHP [22]

Utilizing a markerless motion capture system, this dataset features 8 actors (4 female, 4 male) enacting 8 action sets each. The actions span from simple walks to dynamic activities. It also emphasizes clothing variability and even offers chroma-key masks for background variations.

3) The MuPoTS-3D [62]

this dataset represents a test set designed for multi-person 3D pose estimation. Its authentic 3D poses originate from a markerless MoCap system that encompasses 20 diverse scenes (comprising 5 indoor and 15 outdoor settings). The dataset presents inherent challenges, including occlusions, significant shifts in lighting, and occasional lens flares, especially in outdoor recordings. Across these 20 sequences, the dataset amasses over 8,000 frames, captured by 8 different subjects.

TABLE 2. Datasets for 3D HPE.

Dataset	Year	Type	Num of subjects	Capture method	Environment
HumanEva-I [63]	2010	Single	4 subjects, 6 actions	markerless motion capture	Indoor
Human3.6M [61]	2013	Single	11 subjects, 17 actions	markerless motion capture	Indoor
MPI-INF-3DHP [22]	2017	Single	8 subjects, 8 actions	markerless motion capture	Indoor and outdoor
TotalCapture [64]	2017	Single	5 subjects, 5 actions	IMU, and vicon mocap	Indoor
AMASS [64]	2019	Single	300 subjects	markerless motion capture	Indoor and outdoor
NBA2K [65]	2020	Single	27 subjects	game engine	Indoor and outdoor
CMU Panoptic [66]	2016	Multiple	8 subjects	markerless motion capture	Indoor
3DPW [67]	2018	Multiple	7 subjects	IMU, and camera	Outdoor
MuPoTS-3D [62]	2018	Multiple	8 subjects	markerless motion capture	Indoor and outdoor

E. EVALUATION METRICS FOR 3D HPE

We outline several widely utilized metrics, providing detailed configurations according to datasets.

1) MPJPE (MEAN PER JOINT POSITION ERROR)

This is the most widely used metric to evaluate the performance of 3D HPE. MPJPE is calculated by using the Euclidean distance between the estimated 3D joints and the ground truth positions as given by the equation:

$$E_{MPJPE}(f, \mathcal{S}) = \frac{1}{N_{\mathcal{S}}} \sum_{i=1}^{N_{\mathcal{S}}} \|P^f_{\mathcal{S}}(i) - P^f_{gt,\mathcal{S}}(i)\|_2 \qquad (1)$$

where f represents a frame and $\mathcal{S}$ denotes the associated skeleton. $P^f_{\mathcal{S}}(i)$ signifies the estimated position of joint i, and $P^f_{gt,\mathcal{S}}(i)$ is its corresponding ground truth position. Considering all joints, we have $N_{\mathcal{S}}$ = Number of Joints.

The MPJPE values are subsequently averaged across all frames. Furthermore, the derived normalized metrics are referred to as **NMPJPE**. As the orientation remains consistent, this transformation is less restrictive compared to the popularly employed Procrustes alignment, which we denote as **PA-MPJPE**.

2) MPVE (MEAN PER VERTEX ERROR) [68]

quantifies the Euclidean distances between the actual vertices and their predicted counterparts:

$$E_{MPVE} = \frac{1}{N} \sum_{i=1}^{N} \|V_i - V_i^*\|_2, \qquad (2)$$

where N signifies the total number of vertices, V represents the ground truth vertices, and V^* stands for the predicted vertices.

3) 3DPCK

is a 3D extension of the *Percentage of Correct Keypoints (PCK)* metric utilized in 2D HPE evaluations. A joint estimate is deemed correct if the distance between the predicted value and the actual ground truth falls below a specified limit.

V. METHODOLOGY AND ANALYSIS

In this section, we comprehensively compare an array of 2D and 3D HPE methods designed with an emphasis on efficiency. While previous research surveys have predominantly focused on methods with better accuracy results, our analytical framework adopts a more holistic approach, considering accuracy and computational efficiency. This assessment is manifested through two crucial metrics: the number of parameters and the FLOPs. As delineated in Table 3 and Table 4, while the methods have been sequenced based on their accuracy metrics, it remains paramount for readers to peruse the triad of columns on the extreme right to discern an optimal equilibrium between accuracy and efficiency for a given application.

A. EFFICIENT 2D HPE

Table 3 provides a comprehensive comparison of various efficient 2D Human Pose Estimation (HPE) methods, evaluating their performance across single-person and multiple-people scenarios. For single-person HPE, the methods are assessed on the MPII dataset using the PCK metric, while for multiple-person HPE, the evaluation is based on the COCO dataset using the AP metric.

In the realm of 2D single-person efficient HPE, the table features diverse approaches ranging from B-CNN's binarization and hierarchical, parallel, multi-scale strategy, to NAS HPE's use of Neural Architecture Search (NAS) and spatial information correction modules. The methods demonstrate varying degrees of accuracy, with Fast HPE achieving a high 90.8% accuracy using knowledge distillation. In terms of efficiency, Lite-HRNet-30 stands out with a significantly lower FLOP count of 0.42G, while DSPNet maintains a balance between accuracy (86.3%) and computational efficiency with 0.73G FLOPs.

For 2D multiple-people efficient HPE evaluated on the COCO dataset, the approaches again show diversity, with NAS HPE and DSPNet both achieving high accuracy levels of 76.5% and 69.4%, respectively. Here, CSN's method of suppressing unnecessary channels in conjunction with a grouped bottleneck block shows a promising balance between accuracy (69.9%) and computational efficiency (1.08G FLOPs).

Overall, the table highlights the advancements in efficient HPE methods, illustrating a trend toward achieving higher accuracy without compromising computational efficiency. This is pivotal for real-time applications and deployment on resource-constrained platforms. The variety of approaches,

TABLE 3. Efficient 2D HPE methods comparison.

Method	Backbone	Main ideas	Accuracy	Num of Param	FLOPs
2D single person efficient HPE evaluated on MPII using PCK					
B-CNN [34]	Hourglass	Binarization, Hierarchical, Parallel, and Multi-Scale	78.1	6.2M	-
NAS HPE [69]	MobileNetV2	NAS, spatial information correction module	80.9	5M	-
DSPNet [70]	EfficientNet	Deep supervision pyramid architecture	86.3	7.59M	0.73G
UEPDN [71]	PeleeNet	Single-branch pose knowledge distillation framework	87	2.75M	6.2G
MAPD [72]	HRNet	Multi-Angle Pose Distillation	87.6	1M	2G
Lite-HRNet-30 [73]	HRNet	Light high-resolution network, conditional channel weighting	87.89	1.8 M	0.42G
CSN [74]	ResNet	Suppress unnecessary channels, grouped bottleneck block	88.5	3.8M	1.43G
CVC-Net [75]	Hourglass	Residual block with channel attention mechanism	89.3	4.2M	-
Improved BNN [76]	Hourglass	Binarization with distillation	89.5	6.2M	-
Fast HPE [37]	Hourglass	Knowledge distillation	90.8	3M	9G
Light Stacked [77]	Hourglass	Dilated convolution, depthwise separable convolution	91.6	3.9M	-
2D multiple people efficient HPE evaluated on COCO using AP					
Iterative Pruning [78]	HRNet	pruning profiling, iterative pruning via knowledge distillation	68.7	15.0M	-
DSPNet [70]	EfficientNet	Deep supervision pyramid architecture	69.4	7.59M	1.2G
Simple HPE [79]	ResNet50	Global Context (GC) block based on SimpleBaseline [80]	69.6	2.9M	1G
CSN [74]	ResNet	Suppress unnecessary channels, grouped bottleneck block	69.9	3.8M	1.08G
CVC-Net [75]	Hourglass	Residual block with channel attention mechanism	74.1	4.2 M	-
NAS HPE [69]	MobileNetV2	NAS, spatial information correction module	76.5	5M	-

TABLE 4. Efficient 3D HPE methods comparison on Human3.6 dataset using MPJPE.

Method	Backbone	Main ideas	MPJPE[↓]	Num of Param	FLOPs
MoVNect [81]	CNN Regression	Knowledge distillation	97.3	1.03M	1.35M
VNect [82]	CNN Regression	fully convolutional pose formulation	80.5	14.6M	-
3D Mobile [83]	MobileNetV2	MobileNetV2, skip concatenation structure	56.9	2.24M	3.92G
Deciwatch [84]	ResNet50	samples 10% frames, Transformer-based network	53.5	-	0.621G
Attention Mach [85]	ResNet50	Attention mechanism, temporal contexts	45.1	11.25M	-
MotionAGFormer [86]	Transformer	combining transformer and GCNFormer for improved joint relationship understanding	45.1	2.2 M	1.0 G
HDFormer [87]	U-shaped transformer	High-order Direct Transformer efficient and high-order attention-based model	42.6	3.7 M	0.6 G

from binarization and knowledge distillation to innovative architectural designs, suggests a rich field of ongoing research focused on optimizing both the accuracy and efficiency of HPE systems.

B. EFFICIENT 3D HPE

Table 4 compares efficient 3D Human Pose Estimation (HPE) methods on the Human3.6 dataset using MPJPE, where the lower error (mm) is better. HDFormer excels with the lowest error of 42.6, using a U-shaped transformer and high-order attention, with 3.7M parameters and 0.6G FLOPs. MoVNect, though higher in error (97.3mm), is efficient with 1.03M parameters and 1.35M FLOPs due to knowledge distillation. VNect, with a fully convolutional pose, has an error of 80.5mm and 14.6M parameters. 3D Mobile combines MobileNetV2 and skip concatenation for an errors of 56.9mm, with 2.24M parameters and 3.92G FLOPs. Deciwatch achieves 53.5mm errors with 0.621G FLOPs, sampling 10% of frames. MotionAGFormer, integrating transformer and GCNFormer, records 45.1 mmerror, 2.2M parameters, and 1.0G FLOPs. This comparison underscores the trend towards lower error and efficient 3D HPE methods, crucial for real-time, resource-efficient systems.

Overall, this table reflects the current trend in 3D HPE research towards developing methods that not only achieve high accuracy but also maintain computational efficiency. These advancements are crucial for practical applications of 3D HPE, especially in real-time and resource-limited environments. The variety in approaches, from knowledge distillation to attention mechanisms, illustrates the diverse strategies being explored to optimize both the accuracy and efficiency of 3D HPE systems.

VI. CONCLUSION AND FUTURE DIRECTIONS

In the rapidly advancing domain of efficient monocular human pose estimation (HPE) using deep learning, this survey identifies several critical areas for future exploration, with a particular emphasis on enhancing efficiency. Our contributions in highlighting efficient HPE methods are further underscored in the following key research directions:

A. FOCUSING ON EFFICIENT 3D POSE ESTIMATION FROM MONOCULAR IMAGES

While advancements in 2D HPE are noteworthy, the transition to 3D pose estimation from monocular images poses significant challenges, primarily due to the absence of depth data. Our survey emphasizes the need for research into efficient methods that can bridge this gap, possibly through innovative algorithmic solutions that maintain computational efficiency. The development of lightweight yet effective models for 3D pose estimation represents a pivotal area of future research, aligning with the core theme of our survey.

B. ENHANCING MODEL ROBUSTNESS WHILE MAINTAINING EFFICIENCY

Enhancing the robustness of HPE models in diverse environmental conditions, without compromising their efficiency, is crucial. This survey sheds light on the necessity of developing models that not only adapt to challenges like occlusions and variable lighting but also maintain a low computational footprint. Exploring methods that enhance robustness in an efficient manner, such as lightweight architectures with enhanced generalization capabilities, aligns with our focus on efficiency in HPE methods.

C. PRIORITIZING PRIVACY IN EFFICIENT HPE METHODS

As HPE technologies become more prevalent, ensuring privacy in an efficient manner is paramount. This survey brings forth the importance of designing efficient HPE systems that incorporate privacy-preserving mechanisms from the ground up. Developing models that respect user privacy, while retaining computational efficiency, is a critical area that aligns with our survey's emphasis on efficient HPE methods.

In conclusion, these focal areas represent the critical avenues for research in efficient monocular HPE using deep learning. Addressing these challenges is expected to significantly advance the field and expand the practical deployment of HPE technologies in various real-world scenarios.

REFERENCES

[1] X. Yan, L. Zhang, B. Liu, and G. Qu, "A lightweight and fast approach for upper limb range of motion assessment," in *Proc. 21st IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2022, pp. 53–60.

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.

[3] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors*, vol. 16, no. 12, p. 1966, Nov. 2016.

[4] W. Liu, Q. Bao, Y. Sun, and T. Mei, "Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective," *ACM Comput. Surveys*, vol. 55, no. 4, pp. 1–41, Apr. 2023.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[6] A. Scano, R. M. Mira, P. Cerveri, L. Molinari Tosatti, and M. Sacco, "Analysis of upper-limb and trunk kinematic variability: Accuracy and reliability of an RGB-D sensor," *Multimodal Technol. Interact.*, vol. 4, no. 2, p. 14, Apr. 2020.

[7] T. M. Ghazal, M. Kamrul Hasan, S. Norul Huda Abdullah, K. Azmi Abubakkar, and M. A. M. Afifi, "IoMT-enabled fusion-based model to predict posture for smart healthcare systems," *Comput., Mater. Continua*, vol. 71, no. 2, pp. 2579–2597, 2022.

[8] X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey," *Frontiers Robot. AI*, vol. 7, no. 1, p. 71, Jun. 2020.

[9] J. Zou, B. Li, L. Wang, Y. Li, X. Li, R. Lei, and S. Sun, "Intelligent fitness trainer system based on human pose estimation," in *Proc. 5th Int. Conf. Signal Inf. Process., Netw. Comput.*, 2019, pp. 593–599.

[10] H. Jeon, Y. Yoon, and D. Kim, "Lightweight 2D human pose estimation for fitness coaching system," in *Proc. 36th Int. Tech. Conf. Circuits/Syst., Comput. Commun. (ITC-CSCC)*, Jun. 2021, pp. 1–4.

[11] M. Cormier, A. Clepe, A. Specker, and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?" in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 591–601.

[12] A. M. Sharma, K. Venkatesh, and A. Mukerjee, "Human pose estimation in surveillance videos using temporal continuity on static pose," in *Proc. Int. Conf. Image Inf. Process.*, Nov. 2011, pp. 1–6.

[13] Y. Zhao, Z. Yuan, and B. Chen, "Accurate pedestrian detection by human pose regression," *IEEE Trans. Image Process.*, vol. 29, pp. 1591–1605, 2020.

[14] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Comput. Vis. Image Understand.*, vol. 73, no. 3, pp. 428–440, 1999.

[15] L. Wang, W. Hu, and T. Tan, "Recent developments in human motion analysis," *Pattern Recognit.*, vol. 36, no. 3, pp. 585–601, Mar. 2003.

[16] Y. Chen, Y. Tian, and M. He, "Monocular human pose estimation: A survey of deep learning-based methods," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102897.

[17] Q. Dang, J. Yin, B. Wang, and W. Zheng, "Deep learning based 2D human pose estimation: A survey," *Tsinghua Sci. Technol.*, vol. 24, no. 6, pp. 663–676, Dec. 2019.

[18] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[20] A. Singh, S. Agarwal, P. Nagrath, A. Saxena, and N. Thakur, "Human pose estimation using convolutional neural networks," in *Proc. Amity Int. Conf. Artif. Intell. (AICAI)*, Feb. 2019, pp. 946–952.

[21] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[22] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.

[23] M. F. Ghezelghieh, R. Kasturi, and S. Sarkar, "Learning camera viewpoint using CNN to improve 3D body pose estimation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 685–693.

[24] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. 14th Eur. Conf.*, Jun. 2021, pp. 483–499.

[25] T. Xu and W. Takano, "Graph stacked hourglass networks for 3D human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16100–16109.

[26] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[27] S. Yang, Z. Quan, M. Nie, and W. Yang, "TransPose: Keypoint localization via transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11782–11792.

[28] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "TokenPose: Learning keypoint tokens for human pose estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11293–11302.

[29] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "HRFormer: High-resolution vision transformer for dense predict," in *Proc. Conf. Neural Inf. Process. Syst.*, 2021, pp. 7281–7293.

[30] H. Ma, Z. Wang, Y. Chen, D. Kong, L. Chen, X. Liu, X. Yan, H. Tang, and X. Xie, "PPT: Token-pruned pose transformer for monocular and multi-view human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 424–442.

[31] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1954–1963.

[32] K. Lin, L. Wang, and Z. Liu, "Mesh graphormer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12919–12928.

[33] C. Zheng, M. Mendieta, T. Yang, G.-J. Qi, and C. Chen, "FeatER: An efficient network for human reconstruction via feature map-based transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 13945–13954.

[34] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3726–3734.

[35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[36] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.
[37] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3512–3521.
[38] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.
[39] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
[40] Q. Bao, W. Liu, J. Hong, L. Duan, and T. Mei, "Pose-native network architecture search for multi-person human pose estimation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 592–600.
[41] L. Xu, Y. Guan, S. Jin, W. Liu, C. Qian, P. Luo, W. Ouyang, and X. Wang, "ViPNAS: Efficient video pose estimation via neural architecture search," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16067–16076.
[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[43] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5MB model size," 2016, *arXiv:1602.07360*.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–16.
[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[46] L. Zhang, "Intrusion detection systems to secure in-vehicle networks," Ph.D. thesis, Dept. Comput. Inf. Sci., Univ. Michigan, 2023.
[47] S. R. Young, P. Devineni, M. Parsa, J. T. Johnston, B. Kay, R. M. Patton, C. D. Schuman, D. C. Rose, and T. E. Potok, "Evolving energy efficient convolutional neural networks," in *Proc. IEEE Int. Conf. Big Data*, Dec. 2019, pp. 4479–4485.
[48] L. Zhang, X. Yan, and D. Ma, "A binarized neural network approach to accelerate in-vehicle network intrusion detection," *IEEE Access*, vol. 10, pp. 123505–123520, 2022.
[49] J. Johnson, "Rethinking floating point for deep learning," 2018, *arXiv:1811.01721*.
[50] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[51] S. Jin, L. Xu, J. Xu, C. Wang, W. Liu, C. Qian, W. Ouyang, and P. Luo, "Whole-body human pose estimation in the wild," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer, 2020, pp. 196–214.
[52] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and B. Schiele, "PoseTrack: A benchmark for human pose estimation and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5167–5176.
[53] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. Brit. Mach. Vis. Conf.*, 2010, p. 5.
[54] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681.
[55] W. Zhang, M. Zhu, and K. G. Derpanis, "From actemes to action: A strongly-supervised representation for detailed action understanding," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2248–2255.
[56] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and Y. Wang, "AI challenger : A large-scale dataset for going deeper in image understanding," 2017, *arXiv:1711.06475*.
[57] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "CrowdPose: Efficient crowded scenes pose estimation and a new benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10855–10864.
[58] W. Lin, H. Liu, S. Liu, Y. Li, R. Qian, T. Wang, N. Xu, H. Xiong, G.-J. Qi, and N. Sebe, "Human in events: A large-scale benchmark for human-centric video analysis in complex events," 2020, *arXiv:2005.04490*.
[59] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (Almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, Sep. 2012.
[60] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[61] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[62] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt, "Single-shot multi-person 3D pose estimation from monocular RGB," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 120–130.
[63] L. Sigal, A. O. Balan, and M. J. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 4–27, Mar. 2010.
[64] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–13.
[65] L. Zhu, K. Rematas, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Reconstructing NBA players," in *Proc. 16th Eur. Conf.*, 2020, pp. 177–194.
[66] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3334–3342.
[67] T. V. Marcard, R. Henschel, M. J. Black, B. Rosenhahn, and G. Pons-Moll, "Recovering accurate 3D human pose in the wild using IMUs and a moving camera," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 601–617.
[68] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, "Learning to estimate 3D human pose and shape from a single color image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 459–468.
[69] W. Zhang, J. Fang, X. Wang, and W. Liu, "EfficientPose: Efficient human pose estimation with neural architecture search," *Comput. Vis. Media*, vol. 7, no. 3, pp. 335–347, Sep. 2021.
[70] F. Zhong, M. Li, K. Zhang, J. Hu, and L. Liu, "DSPNet: A low computational-cost network for human pose estimation," *Neurocomputing*, vol. 423, pp. 327–335, Jan. 2021.
[71] S. Zhang, B. Qiang, X. Yang, X. Wei, R. Chen, and L. Chen, "Human pose estimation via an ultra-lightweight pose distillation network," *Electronics*, vol. 12, no. 12, p. 2593, Jun. 2023.
[72] H. Li, "Lightweight and effective human pose estimation model based on multi-angle knowledge distillation," *J. Phys. Conf. Ser.*, vol. 2224, no. 1, Apr. 2022, Art. no. 012025.
[73] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-HRNet: A lightweight high-resolution network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10435–10445.
[74] S. Zhou and L. Peng, "Channel sifted model for pose estimation," *Int. J. Speech Technol.*, vol. 53, no. 9, pp. 11373–11388, May 2023.
[75] X. Qin, H. Guo, C. He, and X. Zhang, "Lightweight human pose estimation: CVC-net," *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 17615–17637, May 2022.
[76] A. Bulat, G. Tzimiropoulos, J. Kossaifi, and M. Pantic, "Improved training of binary networks for human pose estimation and image recognition," 2019, *arXiv:1904.05868*.
[77] S.-T. Kim and H. J. Lee, "Lightweight stacked hourglass network for human pose estimation," *Appl. Sci.*, vol. 10, no. 18, p. 6497, Sep. 2020.
[78] S. Choi, W. Choi, Y. Lee, and H. Woo, "Iterative pruning-based model compression for pose estimation on resource-constrained devices," in *Proc. 5th Int. Conf. Mach. Vis. Appl. (ICMVA)*, Feb. 2022, pp. 110–115.
[79] Z. Zhang, J. Tang, and G. Wu, "Simple and lightweight human pose estimation," 2019, *arXiv:1911.10346*.
[80] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.
[81] D.-H. Hwang, S. Kim, N. Monet, H. Koike, and S. Bae, "Lightweight 3D human pose estimation network training using teacher–student learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 468–477.

[82] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "VNect: Real-time 3D human pose estimation with a single RGB camera," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, Aug. 2017.
[83] S. Choi, S. Choi, and C. Kim, "MobileHumanPose: Toward real-time 3D human pose estimation in mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2328–2338.
[84] A. Zeng, X. Ju, L. Yang, R. Gao, X. Zhu, B. Dai, and Q. Xu, "Deciwatch: A simple baseline for 10× efficient 2D and 3D pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 607–624.
[85] R. Liu, J. Shen, H. Wang, C. Chen, S.-C. Cheung, and V. Asari, "Attention mechanism exploits temporal contexts: Real-time 3D human pose reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5063–5072.
[86] S. Mehraban, V. Adeli, and B. Taati, "MotionAGFormer: Enhancing 3D human pose estimation with a transformer-GCNFormer network," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6920–6930.
[87] H. Chen, J.-Y. He, W. Xiang, Z.-Q. Cheng, W. Liu, H. Liu, B. Luo, Y. Geng, and X. Xie, "HDFormer: High-order directed transformer for 3D human pose estimation," 2023, *arXiv:2302.01825*.

XUKE YAN (Member, IEEE) received the B.S. degree from the Department of Electrical Engineering, University of Electronic Science and Technology of China, in 2014, and the M.S. degree from the Department of Electrical and Computer Engineering, University of Michigan–Dearborn, in 2015. He is currently pursuing the Ph.D. degree with the Computer Science and Engineering Department, Oakland University, USA. His research interests include machine learning, computer vision, and embedded systems.

BO LIU (Senior Member, IEEE) received the B.S. degree from the Department of Automation, Beijing Institute of Technology, Beijing, China, and the M.S. and Ph.D. degrees from the Department of Automation, System Integration Institute, Tsinghua University, Beijing, in 2003 and 2008, respectively. After graduation, she was with NEC Laboratories, China, The University of Chicago, Argonne National Laboratory, Beijing University of Technology, and Massey University. Her research interests include big data, data mining, machine learning, bioinformatics, scientific workflow, semantic web, and ontology reasoning.

GUANGZHI QU (Senior Member, IEEE) received the B.E. and M.E. degrees from the Department of Computer Science and Engineering, Beihang University, and the Ph.D. degree in computer engineering from The University of Arizona, in 2005. He joined the Computer Science and Engineering Department, Oakland University, in 2007, where he currently holds the position of a Professor. His research interests include applied machine learning, artificial intelligence, operating systems, and cybersecurity. Additionally, he served as the Conference Program Co-Chair for the International Conference on Machine Learning and Applications (ICMLA) in 2014 and 2021.