State of the Art in Real-Time 2D Pose Estimation using Smartphones

Jánosi Csongor
it255501@fhstp.ac.at
University of Applied Sciences St. Pölten
St. Pölten, Austria

Abstract

Real-time 2D human pose estimation using smartphone cameras has become a highly relevant technology for interactive applications such as physical education, gesture-based interaction, and augmented reality. While recent advances have significantly improved pose estimation models, using such systems on mobile devices remains challenging due to computational, energy, and latency constraints. This paper presents a state-of-the-art review of real-time 2D pose estimation approaches targeting smartphones, following a PRISMA inspired systematic literature review. Two recent and relevant papers, LMFormer and LiPE, are analyzed and compared based on design choices, efficiency metrics, evaluation scope, and suitability for mobile and human–computer interaction (HCI) practical use cases. The review highlights the balancing of accuracy and computational efficiency, identifies limitations in current evaluation practices, and outlines open research gaps related to end-to-end mobile performance and interaction-oriented evaluation criteria.

CCS Concepts

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools.*

Keywords

state of the art, systematic review, mobile device, pose estimation, 2D pose estimation, skeleton estimation, keypoint detection

ACM Reference Format:

Jánosi Csongor. 2026. State of the Art in Real-Time 2D Pose Estimation using Smartphones. In . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Real-time 2D human pose estimation is being adapted to a wide range of interactive systems, including physical education, gesture-based user interfaces, and augmented and virtual reality applications. Smartphones being a part of our everyday life, there are new opportunities of using pose estimation directly on mobile devices. Many state-of-the-art methods come with high computational costs and are meant for desktop use cases. Smartphones introduce constraints such as limited processing power. The most important goals addressed by the papers below are limited processing power and the need for low-latency feedback. These constraints are highly relevant in human–computer interaction (HCI) contexts, where responsiveness and robustness are often more important than benchmark accuracy.

This SOTA review aims to evaluate lightweight 2D pose estimation methods for smartphone deployment, focusing on design trade-offs, efficiency, and interaction-oriented application potential. This paper does not seek to explore 3D multi perspective pose estimation, or non-mobile systems.

This paper presents a state-of-the-art review of real-time 2D pose estimation methods specifically targeting smartphone-based and resource-constrained devices. Using a systematic PRISMA inspired review process, some of the most recent research papers are analyzed with a focus on efficiency, evaluation methods, and feasibility to interaction-oriented use cases. This review aims to clarify the challenges that remain for practical, real-time pose estimation on smartphones.

2 Methodology

The method for searching academic databases, and for conducting a systematic review of sources found throughout the initial identification of potential sources for the SOTA is described in the following chapter.

2.1 Academic Databases Searched

Two academic digital libraries are used to gather research on the state of the art IEEE Xplore and ScienceDirect. These libraries were chosen due to the topic falling into the Computer Science category which they are relevant for. The use of additional libraries would broaden the collection of research papers in the study, though due to the limited scope the above two were chosen. For consistency, a predefined list of keywords was used across the libraries when searching for research papers. Keywords and advanced search details of the exact searches can be found in Appendix A.

2.2 Systematic Review

Subsequently, a systematic review of the research found in the above libraries is done using PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) adapted and simplified to fit the scope of this paper. PRISMA is a guideline and method for the reporting of systematic reviews, used to ensure careful planning and detailed documentation of the review process. This process is broken down into the following four steps Identification, Screening, Eligibility, and Inclusion discussed in groups below.[3, 4]

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.2.1 Identification. The first step of PRISMA starts with identifying the potential research papers that could end up being included. This step intertwines with the previously mentioned keyword search. It filters the results for studies that in this review's search criteria were published between 2020 and 2026, collecting research papers solely based on the search criteria before anything is screened by manually reading its contents. Duplicate entries were also checked (see Appendix A). At the end of this step the researcher is left with a number of unique papers that fit the search criteria used. [3, 4]In the present review no duplicates were found so the identification phase contains a total of 107 papers.

2.2.2 Screening, Eligibility, and Inclusion. The remaining unique Identified papers are screened using the two middle PRISMA steps out of the four. Screening where the title and the abstract of the papers are read, and Eligibility where the full text is read. Papers are removed that do not align with the topic and that do not fit the scope of the SOTA paper. In the present review the title and abstract screening excluded 102 papers, leaving 5 for full-text eligibility assessment.

After the remaining research papers' full text is read the number of papers that remain is greatly reduced and are considered suitable for inclusion. Of the 5, one study was removed following the full-text read, and 2 papers were removed following the limitation of publication date to 2024-2026, due to the rapid rate of development in the field.

Ultimately two studies met all inclusion criteria for the final SOTA set. Papers that meet all criteria are included in the final SOTA set, then they are used in the State of the Art chapter [3, 4]. Finally LMFormer (Li et al., 2024) and LiPE (Li & Duan, 2025) were selected for their high relevance to the topic. It is important to note that this review does not cover all 2D pose estimation literature.

3 State of the Art

This section summarizes and groups the main existing approaches, models, or systems identified in the literature. The research and development conducted by the papers below enables more and more use cases to implement pose estimation features. Areas such as Human Computer Interactions (HCI), Augmented and Virtual Reality (AR/VR), for features such as gesture-based interactions and assessing physical education and movement analysis.

3.1 LMFormer (Li et al., 2024)

LMFormer is a lightweight CNN–Transformer model designed for 2D human pose estimation. It builds on well-established concepts from computer vision and deep learning by combining convolutional neural networks (CNNs), traditionally used for image processing, with simplified adaptations of transformer architectures originally developed for natural language processing. This hybrid approach enables the aggregation of global body information while maintaining low computational cost, which is essential for practical deployment on mobile devices.

3.1.1 Approach. One of the main goals of LMFormer is to improve efficiency compared to previous pose estimation methods. This is achieved through a multi-feature perspective Token Mixer, which efficiently combines spatial and visual feature information across the image. The role of the Token Mixer is to share information between distant regions of the image, allowing the model to better capture relationships between body parts that are far apart, rather than focusing only on local pixel neighborhoods as in standard CNN-based approaches. The multi-feature perspective further integrates both spatial and channel-level information, enabling the model to consider not only the position of body parts but also the type of visual features they represent [1]. These design choices increase the feasibility of real-time responsiveness, robustness to partial visibility, and whole-body pose understanding, which are important properties for interaction-oriented and mobile applications.

3.1.2 Datasets, Metrics. To validate the effectiveness of LMFormer several benchmark datasets were used including COCO, MPII, and CrowdPose. Common Objects in Context (COCO) contains large-scale real-world images with 250,000 instances of individuals, annotated with 17 keypoints. LMFormer uses the train2017 COCO dataset for training, with its validation(val2017) and development(test-dev2017) splits for measurements. The dataset evaluates results based on several metrics such as Object Keypoint Similarity (OKS), which compares measured keypoint to ground truth keypoints. COCO also measure average Precision (AP) and Average Recall (AR) of keypoints. The MPII (Max Planck Institute for Informatics) Human Pose Dataset like COCO also provides a diverse range of real-world scenarios and human poses. This dataset contains 40,000 human pose samples annotated with 16 key points each. MPII Human Pose Dataset utilizes the PCKh (head-normalized accuracy) evaluation metric, that measures the percentage of correctly localized keypoints based on normalized distance between detected keypoints compared to ground truth values. The third and last dataset used is CrowdPose, that specializes in images with crowded settings. The CrowdPose dataset contains 20,000 images with annotated keypoints similarly to previous datasets and it utilizes the AP evaluation metric known from the COCO dataset. Due to the low computational cost focus an additional evaluation metric Giga Floating-Point Operations Per Second or GFLOPS a standard benchmark for indication of computational complexity [1].

3.1.3 Results, Strengths and Limitations. The results produced by LMFormer on the previously listed datasets are as follows. LMFormer outperforms other lightweight pose estimators such as MobileNetV2 and ShuffleNetV2 in many areas like using significantly fewer parameters and less computational power while still maintaining performance. It performs well in crowded scenes, occluded body parts and generally maintains accuracy despite reduced complexity. On the COCO val2017 dataset LMFormer-B achieves 65.8 AP with 1.9 M parameters and 0.7 GFLOPs, that compared to MobileNetV2 is an improvement for LMFormer-B by approximately 1.0 AP, LMFormer-B also uses $\tilde{2}$0% of the parameters of MobileNetV2 at $\tilde{4}$4% of the GFLOPs. When tested on MPII dataset LMFormer-L reaches 87.6 PCKh with 85% fewer parameters than HRNet-W32. Finally on CrowdPose LMFormer-L achieves 61.0 AP LMFormer also shows improved values in comparison to ShuffleNetV2 and other models listed with all values in the LMFormer paper [1]. LMFormer's strengths are shown in the combination of detail extraction with global body structure awareness on images while maintaining a balance between accuracy and computational costs.

At the same time limitations of the paper can also be observed. There is no mention of direct benchmarking in real-world scenarios with exact devices allowing space for expanding on the models capabilities [1].

3.1.4 Future Work. Future work can be conducted as also explored by the authors. Other than technical improvements to further enhance evaluation metric results, LMFormer has the potential to be adapted for other use cases such as hand or facial keypoint detection and potentially also for animal pose estimation. The further specification for hand and facial keypoints could play a key role in the practical use of LMFormer in mobile device applications for features such as gesture-based interactions and AR based interfaces [1].

3.2 LiPE (Li & Duan, 2025)

LiPE stands for Lightweight Pose Estimator, its design targets efficiency in real-world applications for mobile and resource-limited devices. Its priority is to minimize computational complexity, measured using Multiply-Accumulate Operations (MACs), for image processing to achieve lowered latency, energy use and real-time response.

3.2.1 Approach. To achieve its goals LiPE utilizes a MobileNetV2 backbone, followed by a lightweight upsampling module, and its prediction head. MobileNetV2 is a highly popular image recognition model utilizing CNNs similarly to LMFormer where it is also referred to. LiPE uses MobileNetV2 as its main part or backbone to extract visual features from the input image. LiPE also uses a Squeeze and Excitation (SE) module to dynamically weight more informative feature channels positively while suppresses less relevant ones. Since feature extraction is done at low resolution for efficiency, upsampling is used to restore spatial detail for keypoint localization. A lightweight prediction head, the final part of the model is used to create keypoint heatmaps and final joint locations resulting in the estimated pose. This design enables LiPE to sacrifice minimal accuracy for major efficiency gains [2].

3.2.2 Datasets, Metrics. Experiments are conducted using LiPE on benchmark datasets such as MPII previously discussed as part of LMFormer [1]. The MPII dataset due to its sports and physical activity contents aligns well with LiPE's motivation of assessing physical education and movement analysis. Experiments conducted on LiPE though slightly narrower compared to LMFormer, use a subset of the training data for training the model and validate it on a held-out validation set of 2958 samples. LiPE's performance is analyzed through several evaluation metrics. PCKh an evaluation metric related to the MPII dataset is used to measure performance. The measure of efficiency used by LiPE MACs that are the main target to achieve increased efficiency. As an additional standard metric in the field the number of parameters are also measured. LiPE uses an application specific metric for measuring similarity between estimated and reference pose called Normalized Pose Distance (NPD) [2].

3.2.3 Results, Strengths and Limitations. LiPE demonstrates achieving its goals when showing its effectiveness through results. As comparison baseline models are used ResNet-50 referred to as baseline and MobileNetV2 version using Depthwise Separable Deconvolution (DSDC) like LiPE does as Model 2. LiPE achieved 85.7 PCKh on MPII with 0.81 G MACs and 2.08 M parameters. Compared to the baseline LiPE achieved a reduction of 93.2% MACs and 93.9% in parameters with a sacrifice of 3.2% in PCKh. In contrast to Model 2 which is the most similar to LiPE, LiPE was able to show improved accuracy by 0.8 PCKh with only adding 0.03 M parameters using an equal number of MACs [2].

Strengths are brought out by the results in showing comparable accuracy to other models at a significant efficiency increase. The intended design of the model has proven to achieve the goals of reaching lowered computational costs. While still limited in some areas that show room for improvement. Validation could be improved beyond only using a single dataset for evaluation like LMFormer. Furthermore NPD is formally defined and integrated into a real-time pose estimation and evaluation system, the paper does not report quantitative experimental results based on NPD, focusing instead on standard pose estimation accuracy metrics such as PCKh. Finally it is also mentioned that accuracy of pose estimation drops in cases with severe occlusion [2]

3.2.4 Future Work. Although the LiPE paper doesn't go into detail on future work it mentions the authors hopes for the reduction of occlusions' impact on pose estimators' results. In addition to that the addition of features and developments stages can be further improved. Keypoint-level data augmentation could be applied to aid with occlusions, and larger and more diverse datasets could be used in the training stages.

4 Comparison and Gap Analysis

Now that a selection of papers representing the state of the art (SOTA) in the scope of this paper were introduced. Its important to analyze the collective SOTA presented by LMFormer [1] and LiPE [2] by comparing their mobile-oriented 2D pose estimation approaches. Furthermore highlight design traits and identify gaps.

4.1 Side-by-side comparison: LMFormer vs. LiPE

4.1.1 Design goals and intended use. Both LMFormer and LiPE prioritize efficiency making it possible for the models to be used on resource-limited devices such as smartphones. LMFormer focuses on improving global body understanding using a Token Mixer and its evaluated using established benchmark datasets (COCO, MPII, CrowdPose) [1]. LiPE is even closer aligned with the motivation of this SOTA paper explicitly stating its target to be mobile devices, focusing on sacrifice minimal accuracy for major efficiency gains [2].

4.1.2 Architectural strategies. LMFormer focuses on the use of a multi-feature perspective Token Mixer for collecting global body data to achieve [1]. LiPE emphasizes a simple pipeline depending on a MobileNetV2 backbone with lightweight upsampling and prediction head [2]. LMFormer and LiPE both use similar prediction head, a 1x1 convolution though LiPE adds an SE component to re-weight channels [1, 2].

4.1.3 Evaluation scope and metrics. LMFormer reports multiple accuracy measures based on COCO and MPII datasets, and uses

GFLOPs as measure of efficiency [1]. LiPE makes use of MACs as an efficiency measure and evaluates accuracy through the MPII datasets PCKh value [2].

4.1.4 Effectiveness, Results. LMFormer reports results of multiple variants of its models separated by size. LMFormer-B reports 65.8 AP on COCO val dataset and Surpassing MobileNetV2 among others with requiring less GFLOPs. LMFormer-B also reports a 86.7 PCKh accuracy on the MPII validation dataset using 1.9 M parameters [1]. On the MPII dataset LiPE had produced a 85.7 PCKh using 2.08 M parameters [2]. LMFormer shows accuracy and efficiency on benchmark datasets, while LiPE offers a significant reduction in computational requirements and links this to a concrete user related evaluation workflow for physical education use cases.

4.1.5 Strengths that imply mobile device applications. Real-time response capabilities, both papers aim to produce lightweight models which support low latency. For interactive applications such as gesture recognition and AR overlays. Robust pose understanding LMFormer aims to achieve robustness through understanding global body features that can help with imperfect framing and motion. LiPE add SE weights to try and highlight features that are more important than others to increase its key point detection confidence with a lightweight mechanism. LiPE in addition to developing its model also frames the application in a real-world scenario showing motivation and intent for user oriented feedback loops by using NPD for estimation comparison to a baseline [1, 2].

4.2 Gap analysis

This sub-chapter aims to analyze what is still missing for Real-Time 2D Pose Estimation using Smartphones. The first gap that can be identified is the lack of end-to-end benchmarks using mobile devices. Both papers focus large attention on computational efficiency by measuring GFLOPs and MACs, though benchmarks conducted on popular datasets don't always covert to real-word results [1, 2]. It is clear that more reporting is needed in future work on device specific evaluation metrics such as power usage and latency under realistic camera conditions.

As stated in LiPE common real-word obstacles such as occlusions can still significantly impact results, which requires improvements in robustness by potentially using larger datasets or other features to improve the models.

For model metrics to be translated to HCI outcomes additional testing must be included in research. Both papers focus on metrics (AP, PCKh, MACs, number of parameters). LiPE proposes NPD and a user feedback scenario yet it still doesn't produce data using its evaluation metric [2]. Outcomes originating from user-facing tests could further enhance the qualitative value of the papers with information on responsiveness, usefulness, and errors.

5 Conclusion and Future Work

This paper reviewed the state of the art in real-time 2D pose estimation with a specific focus on smartphone-based and resource-constrained devices. Through a systematic literature review, two representative approaches, LMFormer and LiPE, were analyzed and compared in terms of architecture, efficiency, evaluation scope, and relevance to mobile and HCI-oriented applications. The analysis demonstrates that recent lightweight models are able to achieve competitive pose estimation accuracy while significantly reducing computational complexity, making real-time operation on mobile devices increasingly feasible.

The review shows several gaps where solutions rely on efficiency metrics such as GFLOPs or MACs and benchmark datasets, while end-to-end performance on real mobile device related, interaction, responsiveness, latency and other metrics are not analyzed. LiPE mentions application-level evaluation through its NPD metric; however, no qualitative user studies are conducted. Future research should therefore focus on standardized end-to-end evaluation protocols.

This SOTA paper too has strengths and limitations. Thanks to the systematic review conducted during the research process reproducibility is enhanced. However due to the size limitation papers that could have added to the value of the paper may have been missed from additional databases or excluded.

A Appendix: Search Strategy and Literature Management

A.1 Boolean Search String

For searching academic digital libraries a now common advanced search feature can be used to focus a search on specific keywords, publication date range and many other categories that can be used to narrow down the search results to researcher's predefined criteria. The keywords used in the present review are grouped into a primary topic related set and a device related one.

Keywords of the primary topic:

- 2D pose estimation
- keypoint detection

Device-related keywords:

- Mobile
- smartphone
- real-time

A general Boolean statement that is later adapted for use in the advanced search of the libraries can be seen below.

```
("pose estimation" OR "skeleton estimation"
OR "keypoint detection")
AND ("2D" OR "2 dimensional" OR "two dimensional"
OR "single camera" OR "one camera")
AND ("real-time" OR "live")
AND ("mobile" OR "smartphone")
AND NOT ("3D" OR "3 dimensional" OR "three dimensional"
OR "dual camera" OR "multiple camera")
```

The search was restricted to publications between 2020 and 2026 and to English-language journal or conference articles in the field of Computer Science. The search identified 22 papers from IEEE Xplore and 85 from ScienceDirect.

A.2 Database Access & Reference Management

For accessing the libraries the proxy provided by USTP is used. For the collection of sources, local organization, and finding duplicates Zotero is used.

References

[1] Biao Li, Shoufeng Tang, and Wenyi Li. 2024. LMFormer: Lightweight and Multi-Feature Perspective via Transformer for Human Pose Estimation. *Neurocomputing* 594 (2024), 127884. doi:10.1016/j.neucom.2024.127884

[2] Chengxiu Li and Ni Duan. 2025. LiPE: Lightweight Human Pose Estimator for Mobile Applications towards Automated Pose Analysis. *Cognitive Robotics* 5 (2025), 26–36. doi:10.1016/j.cogr.2024.11.005

[3] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372 (March 2021), n71. doi:10.1136/bmj.n71

[4] Matthew J. Page, David Moher, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and Joanne E. McKenzie. 2021. PRISMA 2020 Explanation and Elaboration: Updated Guidance and Exemplars for Reporting Systematic Reviews. *BMJ* 372 (March 2021), n160. doi:10.1136/bmj.n160