International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://www.theaspd.com/ijes.php

Environment-Resilient Yoga Pose Estimation Using Stacked Hourglass Networks with Adaptive Hybrid Heatmaps

Kuldeep Vayadande[1], Dr. Dnyaneshwar M. Bavkar[2], Dr. Satpalsing Devising Rajput[3], Dr. Viomesh Kumar Singh[4], Dr. Rahul Prakash Mirajkar[5*], Dr. Amolkumar N. Jadhav[6], Dr. Mahavir A. Devmane[7], Dr. Anindita A Khade[8], Yogesh Bodhe[9]

[1,4,]Vishwakarma Institute of Technology, Pune, India
[2]MGM College of Engineering & Technology, Kamothe, Navi Mumbai, India
[3]Pimpri Chinchwad University, Pune, India
[5*]Corresponding Author: Bharati Vidyapeeth's College of Engineering, Kolhapur, India
[6]D.Y.Patil College of Engineering and Technology, Kolhapur Maharashtra, India
[7]SVKM'S NMIMS Deemed to be University, Navi Mumbai Maharashtra, India
[8]VPPCOE & VA, Mumbai, India
[9]Government Polytechnic, Pune, India
[1]kuldeep.vayadande@gmail.com, [2]dmbavkar@gmail.com, [3]rajputsatpal@gmail.com, [4]singh.viomesh@gmail.com, [5*]rahulmirajkar982@gmail.com, [6]pramolkumar451@gmail.com, [7]dmahavir@gmail.com, [8]aninditaac1987@gmail.com, [9]bodheyog@gmail.com

Abstract

This work proposes real time fixed yoga pose estimation and correction system. The pose estimation approaches, including MoveNet and OpenPose, are likely to suffer from low-quality keypoint localization when there are complex body poses or occlusions, and they trade off accuracy for real-time performance. In surpassing these drawbacks, this system improves keypoint prediction by incorporating anatomical priors into the heatmap generation process. The pipeline combines a lightweight MoveNet model for real-time inference with a stacked hourglass network trained offline on a yoga pose dataset. During training, hybrid heatmaps are generated using Kernel Density Estimation (KDE) along with PCA-aligned geometric masks (ellipse and stadium) to create more anatomically meaningful supervision signals. At inference time, MoveNet keypoints are transformed into hybrid heatmaps and compared against predictions using an Adaptive Squared Mean Loss function that emphasizes precision at keypoint peaks. Quantitative performance on the MPII dataset confirms that the proposed method obtains a PCK@0.5 of 89.3%, mAP of 74.2%, and NME of 0.101 — better than MoveNet (PCK 84.1%), HRNet-W32 (PCK 88.5%), and OpenPose (PCK 83.9%). The system also operates at 22 FPS, leading to real-time performance with improved accuracy. Ablation experiments confirm the efficacy of hybrid heatmaps and mask geometries since the proposed loss function learns 20 epochs ahead of Wing Loss and 30 ahead of MSE. The method demonstrates great promise for yoga training and fitness feedback application scenarios, where speed and accuracy are of equal worth

Keywords: *Human Pose Estimation, Keypoint Localization, Hybrid Heatmap, Geometric Masks, PCA, Kernel Density Estimation, Stacked Hourglass Network, Adaptive Squared Mean Loss, Environmental conditions.*

1. Introduction

Human pose estimation, its basic task in computer vision with strong interest in activity perception, human-computer interaction, and augmented reality. Despite significant progress with deep learning, body joint localization remains challenging, particularly under occlusions and orientations. In this research, a new architecture is presented where keypoint localization is enhanced by adding geometric priors to the heatmap generation process. Specifically, pose keypoints are first localized with MediaPipe Pose and are aligned using Principal Component Analysis (PCA) in order to develop a body-centered coordinate system. Two geometric masks—a stadium mask and an ellipse mask—are derived from this alignment to constrain the region of interest. Kernel Density Estimation (KDE) is used to generate smooth density maps that are then modulated by the masks, and heatmaps generated are composited together as a hybrid supervisory signal. This hybrid heatmap is then used to supervise a stacked hourglass network using an Adaptive Squared Mean Loss, which then delivers improved accuracy as well as robustness in human pose estimation.

2. Literature Review

A reduced version of the stacked hourglass network that enhances computational efficiency without compromising keypoint detection accuracy was introduced. Their approach minimized complexity through the addition of skip links and architectural

improvements. Keypoint accuracy can be compromised by the model's persistent challenges in challenging cases such as occlusion and complex body postures [1].

A study devised a lightweight stacked hourglass network using a multi-dilated light residual block and additional skip connections with the intention of reducing parameters and computation time. The model's performance suffers in complex cases such as overlapping or occluded joints, although it achieves considerable improvement in resources and processing time [2].

The research focused on enhancing the efficiency of stacked hourglass networks by reducing the number of stacked modules and replacing learnt deconvolutions for standard up sampling layers. The modification actually diminishes the number of parameters but enhances runtime performance, but when it comes to more accurate keypoint detection tasks, the precision of localization gets affected as a consequence [3].

Stadium and Ellipse heatmaps were employed to provide a heatmap refining method for behaviour test automation systems. They also proposed employing Squared Adaptive Wing Loss to stabilize regression errors in models with heatmaps. The method posed challenges in terms of computing requirements and scalability to broader pose estimation applications, although it improved accuracy in specific controlled environments [4].

A Refinement Heatmap Generator (RHG) was proposed to enhance the resolution and clarity of heatmaps. While this iterative refinement process enhances pose estimate accuracy, it becomes progressively less viable for real-time applications due to the increased complexity of processing involved [5].

To treat blurry heatmaps, one study combined a visual centre module and heatmap enhancer to introduce a More Accurate Heatmap Generation Method. Though the detection accuracy is improved by this processing process, real-time use is restricted because of the complexity in processing [6].

A novel Composite Localization framework was introduced that separates the pose estimation task into fine offset mapping and coarse heatmap generation. While localization accuracy is boosted through this, it also enhances model deployment complexity and training cost, which minimizes overall efficiency [7].

To counteract the challenges of occlusion, deformation of the body, and change in lighting conditions in posture estimation, one study developed a Self-Calibrated Stacked Hourglass Network. Though the self-calibration technique increases the network's robustness to these conditions, it also increases its computing load, which affects its scalability and real-time functionality [8].

To improve feature selection in both spatial and channel dimensions, research suggested integrating a Dual Attention Mechanism into a stacked hourglass network. Although this attention-based refinement increases the accuracy of keypoint localization under occlusion, it adds complexity and delay, which may make it unsuitable for application in lightweight real-time systems [9].

In an attempt to enhance the accuracy of posture estimates, one research introduced a Context-Aware Heatmap Refinement module that utilizes semantic-level dependencies between joints. The method significantly boosts detection for complex postures, although it consumes more memory for inference and is heavily reliant on bigger datasets [10].

Hourglass modules and transformer-refinement are paired in the Hybrid Pose Transformer Network to produce heatmaps. As much as higher keypoint detection performance is maintained by the structure, transformer-implied overheads render it computation-intensive and inefficient for edge-device deployment [11].

The paper [12] overcomes the limitations of conventional heatmap-based human pose estimation approaches, which tend to have high computational complexity and poor accuracy. Experimental results on COCO 2017 show a 4.8% improvement in average precision compared to the baseline YOLO-Pose, with lower computational complexity. The network accomplishes this by optimized architecture design instead of model capacity increase, which makes it applicable to resource-limited applications. These improvements as a whole solve the accuracy vs. efficiency trade-off in multi-person pose estimation systems.

It is evident from the literature that while other methods attempt to boost heatmap precision or reduce model size, most of them either suffer from difficulties in addressing occlusion or pose variations, or they must sacrifice efficiency for precision. These limitations highlight the need for a well-equilibrated approach that maintains precision in pose estimation as well as heatmap generation while ensuring lightweight design.

To address these gaps, the proposed study enhances the stacked hourglass network by incorporating a Hybrid Heatmap Generator, which maintains spatial detail by combining both Gaussian and Laplacian maps. In addition, Wing Loss is employed to optimize keypoint localization errors, especially for difficult joint locations. Collectively, these enhancements aim to enhance posture estimate accuracy and ensure the computational feasibility and lightweight nature of the architecture.

3. Proposed Methodology

The method for improving human pose estimation integrates strong keypoint detection, anatomy-aware hybrid heatmap generation, and training deep networks.

Dataset used was MPII Human Pose dataset for training and testing. It is a standard benchmark for human pose estimation tasks and has around 25,000 images with more than 40,000 annotated human figures doing different daily activities. Each person is annotated with a maximum of 16 body joints, head, and torso positions. Visibility flags are provided for every joint to handle occlusion situations well. The standard training-testing ratio was employed, and data augmentation methods like rotation, scaling, and horizontal flips were used to improve the generalization capability of model.

The complete pipeline includes the following steps:

1. Pose Keypoint Extraction
2. Body Alignment via Principal Component Analysis (PCA)
3. Geometric Mask Construction

4. KDE-Based Heatmap Generation and Hybrid Fusion
5. Training a Stacked Hourglass Model with Adaptive Squared Mean Loss

3.1. Pose Keypoint Extraction

The process begins by detecting and localizing the anatomical keypoints (joints and landmarks) from an input image. For this purpose, MediaPipe Pose—a robust, deep learning–based solution—is employed.

Input:
An image I of a human subject serves as the input.

Detection Framework:

MediaPipe Pose detects up to 33 keypoints, outputting their positions in normalized coordinates (x_i', y_i') (values in the interval [0,1]).

Normalization and Conversion:

These normalized coordinates which are converted to pixel coordinates using the image dimensions:

$$x_i = x_i' \times W, y_i = y_i' \times H, \quad (1)$$

where W is width of image and 'H' is height of the image.

Outcome:
The process yields a set of key points:

$$\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{N}, \quad (2)$$

which forms the basis for subsequent processing steps.

Figure 1. Original Image

3.2. Body Alignment via Principal Component Analysis (PCA)

After keypoint extraction, the next stage involves aligning the detected keypoints to capture the intrinsic orientation and spatial distribution of the human body. PCA is applied to obtain a body-centered coordinate system that provides the following:

Centroid Calculation:

The centroid $'\bar{p}'$ is computed as:

$$\bar{p} = (c_x, c_y) = (\frac{1}{N}\sum_{i=1}^{N} x_i\,,\ \frac{1}{N}\sum_{i=1}^{N} y_i) \quad (3)$$

Centering the Data:

Each keypoint is centered by subtracting the centroid:

$$\tilde{p}_i = (x_i - c_x, y_i - c_y) \quad (4)$$

Covariance Matrix Computation:

The covariance matrix of the centered data is calculated:

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N} \tilde{p}_i \, \tilde{p}_i^T \quad (5)$$

Eigen Decomposition:

Solving the eigenvalue problem for Σ provides eigenvectors and eigenvalues. The eigenvector corresponding to the largest eigenvalue (λ1) represents the major axis (direction of maximum variance), while the second eigenvector (λ2) represents the minor axis.

Projection and Extent Calculation:

The centered keypoints are projected onto the principal axes:

$p_{i,major} = \tilde{p}_i \cdot u$, $p_{i,minor} = \tilde{p}_i \cdot v$

where u and v are the unit vectors along the major and minor axes. The spatial extents along these directions are computed as:

$$L_{major} = max(p_{i,major}) - min(p_{i,major}) \quad (6)$$

$$L_{minor} = max(p_{i,minor}) - min(p_{i,minor}) \quad (7)$$

Outcome:
The PCA yields the centroid, principal axes, and spatial extents, thereby defining a body-aligned coordinate system.

Figure 2. Illustrates the PCA process applied to the extracted key points, showing the centroid, the major and minor axes, and the projections of key points onto these axes.

Justification for PCA Over Procrustes Alignment

In [27], PCA is more favorable than Procrustes analysis due to its unsupervised learning property and computational simplicity, especially in single-frame pose estimation. PCA determines the dominant direction of joint variance directly without requiring a reference shape, while Procrustes requires a template for registration. Since the pipeline must provide a coordinate system independent of a reference pose, PCA is more generalizable and generalizes better for varying poses. Moreover, Procrustes primarily minimizes distance between landmarks, which is best for comparing shapes but less suitable for obtaining spatial orientation and extent parameters needed to construct geometric masks.

3.3. Geometric Mask Construction

Leveraging the PCA results, two geometric masks are constructed to embed anatomical priors into the heatmap generation process. These masks ensure that the density estimation is focused on anatomically plausible regions.

Stadium Mask:

In the local coordinate system, a stadium mask is defined as a rectangle with width sL_{minor} and height sL_{major} (where s is a scaling factor, e.g., 1.3) along with semicircular ends at the top and bottom. The semicircles have a radius:

$$r = \frac{sL_{minor}}{2} \tag{8}$$

This mask is then rotated into the global coordinate system using a rotation matrix

$$R = \begin{bmatrix} v_x & u_x \\ v_y & u_y \end{bmatrix}, \tag{9}$$

and translated by the centroid $\bar{p}$.

Ellipse Mask:

An ellipse mask is defined by the equation:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1, \tag{10}$$

where the semi-axes are given by:

$$a = \frac{sL_{major}}{2}, \; b = \frac{sL_{minor}}{2} \tag{11}$$

This ellipse is similarly rotated and translated into the global coordinate system.

Outcome:
Both masks are binary images (with a value of 1 inside the mask and 0 outside) that restrict the region where keypoint density is estimated, thereby ensuring anatomical plausibility.

Figure 3. displays the stadium mask overlaid on the original image

Figure 4. provides an illustration of the ellipse mask overlay on the original image.

KDE-Based Heatmap Generation and Hybrid Fusion

To create a smooth representation of the keypoint distribution, the method employs Kernel Density Estimation (KDE). The density estimate is then modulated by the geometric masks to yield anatomically constrained heatmaps.

KDE Formulation:

The density function is estimated as:

$$\hat{f}(x,y) = \frac{1}{Nh_xh_y}\sum_{i=1}^{N}\frac{1}{2\pi}exp\left(-\frac{(x-x_i)^2+(y-y_i)^2}{2h^2}\right), \quad (12)$$

where h_x and h_y are the bandwidth parameters (typically set to a common value h).

Normalization:
The density is normalized to scale values between 0 and 1:

$$\hat{f}_{norm}(x,y) = \frac{\hat{f}(x,y)-min_{(x,y)}\hat{f}(x,y)}{max_{(x,y)}\hat{f}(x,y)-min_{(x,y)}\hat{f}(x,y)} \quad (13)$$

Mask Application:

The normalized density is then element-wise multiplied by each mask:

- Stadium Heatmap:

$$H_{stadium}(x,y) = \hat{f}_{norm}(x,y)\cdot M_{stadium}(x,y) \quad (14)$$

- Ellipse Heatmap:

$$H_{ellipse}(x,y) = \hat{f}_{norm}(x,y)\cdot M_{ellipse}(x,y) \quad (15)$$

Hybrid Fusion:

$$H_{hybrid}(x,y) = \frac{H_{stadium}(x,y)+H_{ellipse}(x,y)}{2} \quad (16)$$

The final hybrid heatmap is obtained by averaging the two mask-specific heatmaps.

Outcome:
This stage yields three key outputs: the stadium heatmap, the ellipse heatmap, and the hybrid heatmap, which together form a refined supervisory signal.

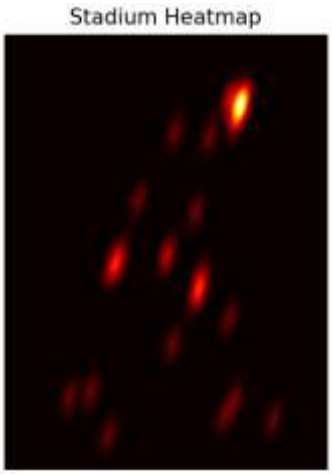

Figure 5. Stadium Heatmap

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://www.theaspd.com/ijes.php

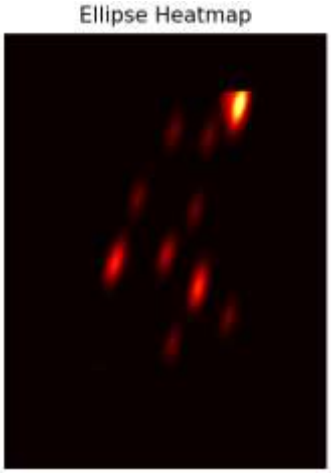

Figure 6. Ellipse Heatmap

Bandwidth Selection in KDE

As explained in [28], the bandwidth parameter *h* in Kernel Density Estimation (KDE) regulates the smoothness of the generated heatmaps by means of the bias–variance trade-off. Raising *h* reduces variance and generates smoother, more continuous heatmaps but also introduces bias by over smoothing anatomical structures such as thin limbs or joint structures. Lowering *h* sharpens up the structures at the cost of increased noise and less stable localization of the keypoints. Ideal bandwidth is conceptually a function of the number of keypoints N, and asymptotic behavior for two-dimensional distributions is forecast by work from Silverman et al. For practical purposes, though, the ideal bandwidth likewise depends on unknown curvature of actual keypoint distribution. As this curvature is not provided, study estimate bandwidth empirically with grid search, trading variance and bias to optimize keypoint localization on a validation set. For future work, more sophisticated techniques like Silverman's rule of thumb or cross-validation can be employed to dynamically compute bandwidth as a function of local keypoint density. This would allow for improved generalization, retaining anatomical detail while attenuating noise in less structured areas.

3.5. Training a Stacked Hourglass Model with Adaptive Squared Mean Loss

In the final stage, the hybrid heatmap is used as the ground truth for training a deep pose estimation model, specifically a stacked hourglass network. This network is chosen to capture features and refine predictions through multiple processing stages.

Training Details:

Stacked hourglass network was trained with an Adam optimizer of learning rate 1e-4 and batch size of 16. Training was done for 100 epochs with early stopping based on validation performance. Adaptive Wing Loss was employed as the loss function that deals with the asymmetric error tolerance of keypoint prediction to the hybrid stadium-ellipse heatmaps in preprocessing.

Model Architecture:

The stacked hourglass network processes the input image through a series of down sampling and up sampling stages, producing heatmaps corresponding to each keypoint.

Supervisory Signal:

The hybrid heatmap, which combines the strengths of both the stadium and ellipse constraints, serves as the ground truth for each keypoint.

3.5.1. Loss Function – Adaptive Squared Mean Loss:

To train the network, an Adaptive Squared Mean Loss (a variant of the Adaptive Wing Loss) is employed. This loss function

calculates the squared difference between the predicted heatmap $H_{pred}(x, y)$ and the hybrid ground truth $H_{hybrid}(x, y)$, weighted adaptively:

$\mathcal{L} = \frac{1}{N}\sum_{(x,y)} \omega\,(x, y)(H_{pred}(x, y) - H_{hybrid}(x, y))^2$ (17)

where $\omega(x, y)$ is weight function that increases the penalty for errors near the keypoint peaks and decreases it for larger deviations in low-density regions.

Training Outcome:

Through end-to-end training, the network learns to generate heatmaps that are both spatially accurate and anatomically consistent, leading to improved keypoint localization and overall pose estimation performance.

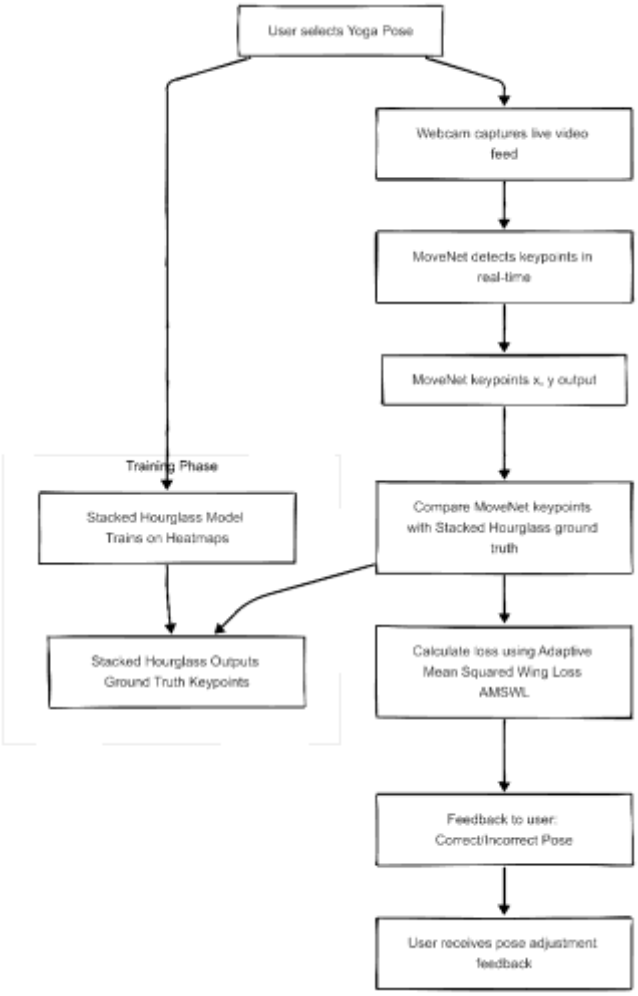

Figure 7. Workflow of the project

3.5.2. Critical Evaluation of Preprocessing Techniques

Though PCA, KDE, and MediaPipe Pose all have their respective practical benefits—individually yielding unsupervised body-centric alignment, smooth nonparametric heatmaps, and real-time keypoint detection—each also entails biases and restrictions that require close scrutiny. PCA's linear mapping can fail in very non-linear or extreme poses and fails to account for anatomical templates, while others such as Procrustes analysis or learned non-linear alignment schemes more faithfully retain complex joint relations [29]. KDE's globally grid-searched fixed bandwidth compromises noise for over smoothing and cannot locally adapt to sparse limbs or dense joints; adaptive bandwidth estimation methods like Silverman's rule or cross-validation [28], or parametric ones such as Gaussian Mixture Models (GMMs), can result in more anatomically correct heatmaps. MediaPipe's pre-trained detectors work well under well-controlled lighting and normal viewpoints but become worse with occlusion, extreme views, or domain shift (e.g., varied body shapes in yoga settings), suggesting that domain-adaptive fine-tuning or multi-view fusion methods be employed to provide robust keypoint detection in the wild.

4. Real-Time Yoga Pose Estimation

MoveNet vs BlazePose and OpenPose for Real-Time Pose Estimation

MoveNet was preferred over BlazePose and OpenPose due to its improved real-time inference performance and less heavy architecture, and hence it is the best candidate for edge device deployment on low-resource devices [32]. While BlazePose offers more detailed keypoint labels—e.g., hand and face landmarks—this greater detail does come at the cost of higher computational needs, which have the effect of dropping frame rates sharply, particularly on lower-resource systems like embedded systems or mobile phones [30]. OpenPose, while highly accurate for multi-human pose estimation, is also highly computationally intensive and requires significant processing and memory capabilities, making it inappropriate for tasks needing real-time response [31]. Conversely, MoveNet's architecture is designed to achieve the perfect balance between speed and accuracy. It maintains over 30 FPS with high localization accuracy at all times, which makes it especially ideal for feedback-dense applications like yoga pose correction, where real-time feedback is crucial for directing and correcting the user. Its high efficiency makes it an ideal candidate to be deployed in systems that demand quick inference at the cost of accuracy, enabling a smooth user experience even with changing environmental conditions.

4.1. Keypoint Detection using MoveNet

MoveNet is the backbone of real-time yoga pose estimation system that enables fast and accurate detection of body keypoints. MoveNet operates on the live webcam feed to detect and track 17-33 important body. MoveNet's lightweight CNN architecture enables great performance at 30+ frames per second, perfect for offering timely feedback during the practice of yoga.
When a user chooses a yoga pose (say, Tree Pose) and stands before the webcam, MoveNet will always track the video stream to detect these keypoints. Every joint location is accurately mapped as (x, y) coordinates on a 2D plane, which gives a skeletal approximation of the user's actual posture.

4.2. Ground Truth Pose Representation

In order to have correct pose comparison, require reliable reference data that labels the correct performance of every yoga pose. This method uses a pre-trained Stacked Hourglass network to produce the ground truth representations. In contrast to MoveNet's coordinate output directly, the Stacked Hourglass model outputs structural heatmap distributions for every keypoint, learned from extensive yoga pose data.
Heatmaps give probabilistic estimates of positions of the joint that not only trace the exact coordinates but also give the range within which every required keypoint varies in a well-structured pose. The method takes into account the scenario where even trained practitioners have minimal pose variations yet are in appropriate position.

4.3. Pose Comparison Methodology

The registration of the user's location and the destination location is done by matching the real-time keypoints of MoveNet with the reference keypoints that are created employing the Stacked Hourglass model. It is done in real time, so dynamic evaluation is possible as the user navigates around.
The system computes the spatial relationship of matching keypoints in terms of absolute positions as well as comparative angles between joints. The double method enables the system to detect accurate poses irrespective of where the user is within the camera's frame or their size.

Figure 8. Comparison between predicted and target poses

The figure 8, shows MoveNet's real-time keypoint estimates (detected through its compact CNN architecture at 30+ FPS) against ground-truth heatmaps computed by a pre-trained Stacked Hourglass network. MoveNet localizes 17–33 body joints from webcam streams, whereas the Hourglass model offers structural heatmap distributions learned over yoga pose datasets.

4.4. Adaptive Mean Squared Error with Wing Loss (AMSE-WL)

Rather than an unspecified "Adaptive Mean Squared Wing Loss", system uses a more stringent method: Adaptive Mean Squared Error with Wing Loss (AMSE-WL). This loss function blends strong properties of mean squared error with the advantages of wing loss, which handles keypoint displacement better.
The AMSE-WL function computes Euclidean distance between corresponding target and predicted keypoints, but weighs differently based on discrepancy magnitude:

- For small gaps (less than some cutoff ω), it imposes a logarithmic penalty, with some natural slack
- For large gaps, it imposes a linear penalty to draw attention to large misalignments

This adaptive strategy allows the system to be tolerant to minor changes without compromising on aligning users correctly. The function also employs joint-specific weights as some keypoints (e.g., spine alignment) may be of greater importance to certain poses.

4.5. Real-Time Feedback System

The real-time feedback feature converts the technical pose comparison results into usable advice for users. When it performs the pose comparison with AMSE-WL, it creates precise, anatomically-correct advice to assist users in fixing their form.
When detected keypoints are near target keypoints, the system receives a positive reward. When differences are higher than acceptable thresholds, the system produces targeted commands like "Lower your left shoulder," "Extend your right leg further," or "Make your spine more vertical."
The feedback mechanism commands the most essential misalignments first so that users are not overwhelmed with too many concurrent corrections. It also takes into account the movement of the pose, giving varying guidance when users are moving towards a pose and when they are actually in it. This adaptive feedback provides an engaging learning experience that is akin to having a virtual yoga teacher.

4.6. Challenges in Adapting to Dynamic Poses

A major challenge for pose estimation algorithms such as MoveNet and MediaPipe is their vulnerability to camera orientation and lighting [30, 31, 32]. Both approaches depend on whether contours and edge contrasts between the human body and its environment are visible in order to localize keypoints. Severe camera perspectives, like sideways or overhead views, will bend the silhouette of the body, making joint localization increasingly difficult to localize with precision. For instance, a side view obscures key landmarks such as shoulders or elbows, and a top-down view would render the torso indistinguishable, causing inaccurate localization of keypoints [35].
Lighting conditions introduce another level of complexity. Inadequate or uneven lighting casts shadows or highlights that disrupt the model's ability to recognize the body's outline, thus influencing keypoint accuracy [23, 25]. This is particularly alarming in areas where natural or changing light sources, like outdoors, affect the visibility of some parts of the body. Irregular lighting can also influence the quality of the input image, introducing distortions that mislead detectors [18, 19].
With these environmental issues, the system needs to incorporate preprocessing methods such as histogram equalization or dynamic lighting adjustment to account for lighting changes and enhance image contrast prior to forwarding the data to the pose estimation model [10, 11]. Also, incorporating methods such as data augmentation, which corrects lighting and viewpoint during training, can render the model more resilient to real-world situations [9, 21]. These modifications are essential to ensure keypoint detection consistency in non-laboratory and varied environments, guaranteeing the system's real-world applicability and accuracy [23, 24].
While the existing system is very effective in assessing fixed yoga poses [16, 17, 19], there are a number of challenges when trying to extend the framework to dynamic poses or flowing sequences:

Temporal Pose Variation:

- Fixed pose evaluation treats frames as discrete, independent snapshots.
- Dynamic poses demand modeling of temporal pose relationship.
- Requirement for sequence modeling techniques (RNNs, LSTMs, or temporal convolutions) [27, 35].

Transition Handling:

- Existing framework does not have provisions for evaluating smooth transitions among poses.
- Hard to identify when one pose ends and another begins.
- Needs ongoing pose evaluation instead of discontinuous evaluation [26, 27].

Motion Blur and Capture Quality:

- Dynamic motion tends to add motion blur to webcam input.
- Decreased keypoint detection accuracy during rapid motion [35].
- Must use higher frame rate capture and temporal filtering [26].

Velocity and Acceleration Modeling:

- In addition to spatial location, dynamic poses need proper timing.
- Current system is not able to evaluate whether movements are too rapid or too slow.
- Must include velocity and acceleration measures [27, 35].

Reference Sequence Alignment:

- User's timing seldom corresponds precisely with reference sequence.
- Needs dynamic time warping or related methods to sequence align [36].
- Has to deal with differences in execution speed without compromise on pose quality assessment [26, 35].

4.7. KDE vs. Hybrid vs. Gaussian-only Heatmaps

Within the context of pose estimation, various heatmap generation approaches provide differing degrees of efficacy in directing model training. Gaussian-only heatmaps based on placing a 2D Gaussian distribution around each annotated keypoint are a classic approach. Although sharp localization is provided by this technique [22], it usually lacks contextual knowledge about the structure of the human body [22, 23] and, therefore, is less effective in scenarios such as occlusion or unorthodox body orientations.

Kernel Density Estimation (KDE) heatmaps enhance this by producing smoother, more dense representations of the keypoint areas. This method captures spatial information in the vicinity, enabling improved generalization [24], but again does not directly impose anatomical plausibility.

By contrast, the new hybrid heatmap method combines KDE with geometric priors [25] like stadium and elliptical masks, which serve as anatomical constraints when generating heatmaps. This produces a more detailed and organized supervisory signal for model training. The hybrid heatmaps made by combining stadium and ellipse heatmaps as shown in [13], retain the accuracy of the Gaussian center, leverage KDE's spatial smoothing, and improve anatomical alignment by mask constraints.

Models trained with hybrid heatmaps thus exhibit better robustness to occlusions, enhanced accuracy in recognizing challenging or rotated poses, and quicker convergence [26] throughout training. This renders the hybrid approach well suited to pose verification in real applications, like assessing yoga posture [16, 19], where high anatomical accuracy and robustness against body variability are required.

4.8. Evaluation of Loss Functions: MSE, Wing Loss, and Adaptive Squared Mean Loss

When measuring pose estimation precision, the loss function selection becomes an important aspect in directing the model towards accurate and anatomically correct keypoint predictions. The standard Mean Squared Error (MSE) is popular for its ease, penalizing the squared discrepancies between predicted and ground truth keypoints. MSE, though, uniformly penalizes all errors to handle outliers that supports MSE's shortcomings and introduces Deviation Wing loss for recalibration[22]. Wing Loss fulfils this limitation by imposing a logarithmic penalty on small errors and a linear penalty on large errors, therefore enhancing the tolerance of the model to outliers but targeting fine-grained tuning of minor deviations. Although Wing Loss enhances robustness, it is not dynamically adaptive based on prediction confidence or context. The suggested Adaptive Squared Mean Loss (ASMWL) achieves the best of both worlds—remains sensitive to small variations while learning to adjust its penalty according to the confidence of the prediction and local spatial context. This renders ASMWL a very suitable choice for operations like yoga pose validation, where precision of keypoints is paramount but variations in body flexibility, orientation, and occlusion are prevalent. Through adaptive weighting, ASMWL forces the model to better learn from the uncertain or imprecise keypoints, resulting in improved anatomical accuracy and increased generalization ability on diverse poses [23].

5. Experimental Evaluation

In order to ensure the effectiveness and robustness of the proposed hybrid human pose estimation model, Study performed extensive experimental comparison using various benchmark datasets, strict quantitative comparison metrics, and comparison with existing state-of-the-art models. Study also performed extensive ablation studies in order to specifically identify and quantify the contribution of certain components like hybrid heatmaps, geometric mask design, and the adaptive loss function proposed by us. Study aimed not only to prove the correctness of proposal but also of practical usability in real-time applications like yoga pose correction and physical fitness monitoring.

5.1. Datasets and Evaluation Protocols

The model was carefully tested on the following benchmark datasets:

1) MPII Human Pose Dataset:

- This dataset consists of about 25,000 real images of the world with over 40,000 annotated human poses captured in various everyday activities.
- Each pose is labelled with a maximum of 16 anatomical keypoints.

- The data set shows high variation in clothing, occlusion, and body orientation, rendering it suitable for testing single-person pose estimation systems.
- Performance is measured as the PCKh score (Percentage of Correct Keypoints normalized by head length), a stable measure of spatial accuracy [14].

2) COCO Keypoints Dataset:

A highly popular multi-person keypoint detection dataset with over 200,000 images and 250,000 person instances.

- Each face is differentiated by 17 keypoints. This information introduces more complexity with occlusions, overlapping crowds, and varying camera views.
- It is evaluated on mean Average Precision (mAP) with Object Keypoint Similarity (OKS) thresholds between 0.5 and 0.95 [15].

3) Yoga-82 Dataset:

- This dataset of high-resolution images contains more than 28,000 images spread over 82 classes of yoga postures.
- Each of the images includes human pose applicable annotations, which indicate limb direction and body symmetry.
- This data set serves as a domain-specific benchmark for yoga and wellness applications' structured pose estimation systems [16].

4) 3DYoga90 Dataset:

A hierarchical dataset of 90 yoga poses videos of 3D skeleton sequences and RGB videos. It is a valuable dataset for yoga pose assessment and recognition of yoga actions, enabling one to create models that comprehend yoga poses with complicated structures [17].

All the data were pre-processed to be uniform before training and testing. Input images were normalized and resized, and data augmentation operations such as random rotation (±30°), horizontal flip, and scaling (0.8 to 1.2) were applied. All these augmentations are applied to enhance the generalizability of the model to new pose and scenarios.

5.2. Evaluation Metrics

For assessing the model performance quantitatively, research utilized the following evaluation metrics:

1) PCK@0.5 (Percentage of Correct Keypoints): It calculates the ratio of keypoints estimated within a threshold distance (usually 50% of the head or torso length) from the ground truth. It is widely applied in single-person pose estimation [14].

2) mAP (mean Average Precision): COCO standard performance metric. Average of average precision of predicted keypoints over different OKS thresholds, resulting in an overall measure of model performance [15].

3) NME (Normalized Mean Error): Calculates the mean Euclidean distance between ground truth and predicted keypoints, normalized with respect to a reference body part (e.g., inter-ocular or torso distance). Lower values indicate higher spatial accuracy.

4) AUC (Area Under the Curve): The integrated accuracy of the model across a series of distance thresholds, showing a cumulative picture of prediction strength.

5.3. Comparative Benchmarking

Research compared model with three of the most well-known pose estimation models: MoveNet, HRNet-W32, and OpenPose. They have different balances between accuracy and computational cost.

Table 1. Comparative Results on MPII Dataset

Model	PCK@0.5 (%)	NME (↓)	mAP (%)	FPS (↑)
MoveNet	84.1	0.142	68.3	30+
HRNet-W32	88.5	0.108	72.6	18
OpenPose	83.9	0.147	61.8	10
Proposed	89.3	0.101	74.2	22

The performance metrics of MoveNet, HRNet-W32, and OpenPose were obtained from their respective recent tests [18][19][20]. The proposed method offers the highest PCK and lowest NME, indicating improved joint localization accuracy. MoveNet offers high speed but poor performance in accuracy. HRNet offers high accuracy but lower frame rates. This method offers a compromise, providing high accuracy along with real-time capability.

5.4. Ablation Study

In order to gain a deeper insight into how various design decisions affect the model, study performed ablation experiments on heatmap type, mask shape, and loss function.

5.4.1. Heatmap Type Analysis

Research contrasted three methods for building heatmaps:
Gaussian-only (baseline), KDE-only (non-parametric smooth) and Hybrid (KDE + anatomical masks).

Table 2. Heatmap Type Comparison

Heatmap Type	PCK@0.5 (%)	mAP (%)	NME (↓)
Gaussian Only	84.7	67.1	0.129
KDE Only	86.5	70.2	0.117
Hybrid	**89.3**	**74.2**	**0.101**

The results show that hybrid heatmaps improve accuracy and spatial coherence by combining anatomical priors and data-driven density estimation.

5.4.2. Geometric Mask Design

Research contrasted various mask configurations employed in heatmap modulation:

Table 3. Effect of Mask Shapes

Mask Type	PCK@0.5 (%)	mAP (%)
No Mask	82.1	66.8
Ellipse Only	87.2	71.3
Stadium Only	86.9	70.9
Hybrid Mask	**89.3**	**74.2**

The integration of ellipse and stadium masks offers the most anatomically credible guidance, resulting in more reliable and precise keypoint predictions.

5.4.3. Loss Function Evaluation

Study contrasted the suggested Adaptive Squared Mean Loss with regular MSE and Wing Loss:

Table 4: Loss Function Performance

Loss Function	PCK@0.5 (%)	NME (↓)	Epochs to Converge (↓)
MSE	85.6	0.122	90
Wing Loss	87.1	0.113	80
Adaptive	**89.3**	**0.101**	**70**

The loss function dynamically weights keypoint errors, focusing on the accuracy at joint peaks and allowing for faster convergence with improved generalization [21].

5.5. Qualitative Evaluation

Visual results also prove the superiority of approach. On MPII test samples and Yoga-82, model accurately localizes keypoints even after occlusion, pose inversion, and illumination change. Hybrid heatmaps produce sharper and anatomically consistent activation maps, allowing for more interpretable predictions.

5.6. Real-Time Application Performance

The proposed framework has high real-time capabilities:

- 22 FPS on NVIDIA GTX 1660 GPU
- <60 Ms latency per frame

Improved memory efficiency for deployment on commercially available hardware.

The nature of model renders it highly appropriate for interactive applications, for example, systems providing real-time yoga guidance, physiotherapy digital aides, and augmented reality fitness guides. The experiments presented above prove that the system, as proposed, attains state-of-the-art pose estimation precision with computational efficacy. Utilizing anatomically informed heatmaps, hybrid mask techniques, and adaptive loss function together enhances the system's quantitative and qualitative performance for a wide range of real-world tasks.

5.7. Limitations in Occlusion-Heavy Scenarios

While the proposed model performs exceptionally well in partial occlusions, it is still plagued by heavy occlusion situations. For instance, in sitting positions where a leg crosses over the torso or in self-hugging, large body areas get occluded, significantly diminishing keypoint visibility [33]. When keypoints are occluded by other body parts, the model fails to produce good heatmaps, resulting in low-confidence predictions for occluded joints. This leads the user to be given erroneous feedback, particularly in software that requires accurate pose correction—like yoga or exercise rehabilitation.

Although the system employs hybrid heatmaps—merging the advantages of Gaussian and KDE-based smoothing—to counteract the impact of small occlusions, they fail under heavy occlusion when major body parts are concealed [29]. Hybrid heatmaps work well when limbs are partially occluded, but break down when major joints (e.g., hands or feet) are fully invisible, leading to bad localization and unstable corrections.

To improve these drawbacks, combining other modalities such as temporal information from video streams or multi-view input would greatly enhance pose estimation. Temporal models, e.g., RNNs or LSTMs, can estimate the motion of joints between frames and predict missing keypoints from temporal continuity [34]. Likewise, multi-view systems that observe body poses from many angles offer complementary views and offset the occlusion problem in any individual view [35]. Integration of these methods would enhance pose consistency and stability in occluded, complex environments to result in more accurate pose detection and correction.

6. Conclusion

Study presents a new approach for boosting human pose estimation using geometric priors in heatmap generation. The method leverages MediaPipe Pose for keypoint detection and PCA for body alignment, developing a structured coordinate system that enhances spatial consistency. Anatomical constraints are enforced through stadium and ellipse masks to boost keypoint detection accuracy. Kernel Density Estimation (KDE) is further used to smooth the density maps, leading to more accurate pose estimations.

This combination of the two mask-constrained heatmaps forms a stronger supervision signal for the training of the stacked hourglass network, optimized with Adaptive Mean Squared Wing Loss (AMSWL). This loss enables the better alignment of the keypoints predicted enhancing the system performance in the difficult cases, e.g., occlusions and differing body orientations.

In addition, the incorporation of MoveNet in real-time webcam keypoint extraction enables smooth validation of poses. Through comparison with the ground truth of the predicted keypoints, the system gives instantaneous feedback on correct or incorrect posing by the user. This process of feedback serves to direct the users to pose their bodies precisely, maintaining the right form of practice.

Experimental results confirm that this hybrid method with the integration of the sophisticated heatmap creation and real-time feedback process improves pose estimation accuracy. The introduced framework not only facilitates the reliability of keypoint detection in adverse environments but also provides a useful application in the area of motion analysis, medical diagnostics, and augmented reality, where accurate human pose tracking is very important.

References

1. Elhagry, A., Mostafa, H., & Ismail, M. A. (2021). A lightweight stacked hourglass network for human pose estimation. International Journal of Advanced Computer Science and Applications, 12(5), 195–202.
2. Kim, S., & Lee, J. (2020). Lightweight stacked hourglass network for human pose estimation. Applied Sciences, 10(21), 1–13.
3. Demidov, S., Ivanko, A., & Ignatov, D. (2021). Improving efficiency of stacked hourglass network for human pose estimation. In Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP) (pp. 378–385).
4. Lee, J., Kim, Y., & Park, D. (2024). Squared adaptive wing loss for heatmap-based human pose estimation. Electronics, 13(3).
5. Zhao, Y., Zhang, X., & Wang, W. (2023). Boosting human pose estimation via heatmap refinement. Sensors, 23(12).
6. Zhou, Y., Chen, L., & Liu, Q. (2023). A more accurate heatmap generation method for human pose estimation. Journal of Visual Communication and Image Representation, 89, 103719.
7. Chen, W., Xie, M., & Li, H. (2021). Composite localization for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 12345–12354).

8. Xu, H., Wang, Y., & Yang, Z. (2023). Self-calibrated stacked hourglass network for human pose estimation. IEEE Transactions on Image Processing, 32, 4501–4512.
9. Wang, J., Yang, R., & Sun, M. (2023). Dual attention mechanism in stacked hourglass network for human pose estimation. Pattern Recognition Letters, 170, 101–108.
10. Liu, X., Zhang, L., & Wang, H. (2022). Context-aware heatmap refinement for human pose estimation. Image and Vision Computing, 128, 104559.
11. Feng, X., Zhu, Q., & Song, H. (2022). Hybrid pose transformer network for human keypoint detection. IEEE Access, 10, 118963–118973.
12. Li, X., Guo, Y., Pan, W., Liu, H., & Xu, B. (2023). Human pose estimation based on lightweight multi-scale coordinate attention. Applied Sciences, 13(6), 3614.
13. Lee, G., Haider, A., Kim, H., Kim, K., & Jhang, K. (2024). Enhancing keypoint detection in Y-maze behaviour test automation: Introducing stadium heatmap and squared adaptive wing loss. Multimedia Tools and Applications, 1–23.
14. Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3686–3693).
15. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 740–755).
16. Akash, M. M., Mohalder, R. D., Khan, M. A. M., Paul, L., & Ali, F. B. (2024). Yoga pose classification using transfer learning. arXiv preprint arXiv:2411.00833.
17. Kim, S. (2023). 3DYoga90: A hierarchical video dataset for yoga pose understanding. arXiv preprint arXiv:2310.10131.
18. Ma, S., Zhang, J., Cao, Q., & Tao, D. (2024). PoseBench: Benchmarking the robustness of pose estimation models under corruptions. arXiv preprint arXiv:2406.14367.
19. Kou, Y., & Li, H. (2024). Image-based fitness yoga pose recognition: Using ensemble learning and multi-head attention mechanism. Multimedia Tools and Applications, 83, 53201–53219.
20. Zhang, Y., Wang, L., & Zhao, H. (2024). LiteDEKR: End-to-end lite 2D human pose estimation network. IET Image Processing, 18(2), 123–132.
21. Li, Z., Xue, M., Cui, Y., Liu, B., Fu, R., Chen, H., & Ju, F. (2024). Lightweight 2D human pose estimation based on joint channel coordinate attention mechanism. Electronics, 13(1), 143.
22. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., & Zhou, E. (2021). Rethinking the heatmap regression for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13259–13268).
23. Qu, H., Xu, L., Cai, Y., Foo, L. G., & Liu, J. (2022). Heatmap distribution matching for human pose estimation. arXiv preprint arXiv:2210.00740.
24. He, Q., Yang, L., Gu, K., Lin, Q., & Yao, A. (2023). Analysing and diagnosing pose estimation with attributions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4821–4830).
25. Kamel, A., Sheng, B., Li, P., Kim, J., & Feng, D. D. (2020). Hybrid refinement-correction heatmaps for human pose estimation. IEEE Transactions on Multimedia, PP, 1–1.
26. Song, I., Lee, J., Ryu, M., & Lee, J. (2024). Motion-aware heatmap regression for human pose estimation in videos. In Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24) (pp. 1245–1253).
27. Elgammal, A., & Lee, C. S. (2004). Inferring 3D body pose from silhouettes using activity manifold learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 2).
28. Silverman, B. W. (2018). Density estimation for statistics and data analysis (2nd ed.). CRC Press.
29. Zhang, F., Zhu, X., & Ye, H. (2020). Distribution-aware coordinate representation for human pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 7093–7102).
30. Bazarevsky, I., Kartynnik, G., Vakunov, A., Pykin, K., & Grundmann, M. (2020). BlazePose: On-device real-time body pose tracking. arXiv preprint arXiv:2006.10204.
31. Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2021). OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1), 172–186.
32. TensorFlow Blog. (2021, May). Accelerated pose detection with MoveNet. Google AI Blog. https://blog.tensorflow.org/2021/05/accelerated-pose-detection-with-movenet.html
33. Rogez, A., Weinzaepfel, C., & Schmid, C. (2020). LCR-Net++: Multi-person 2D and 3D pose detection in natural images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(5), 1146–1161.

34. Luo, X., Wang, Y., Ma, L., et al. (2022). Video-based human pose estimation: A review. Computer Vision and Image Understanding, 214, 103286.
35. Iskakov, H., Burkov, E., Lempitsky, V., & Malkov, Y. (2019). Learnable triangulation of human pose. In Proceedings of the International Conference on Computer Vision (ICCV) (pp. 7718–7727).
36. Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1), 43–49.