PoseDepth-CMP: A Comparative Analysis of OpenPose COCO and MPI Keypoint Models for Contour-Guided Monocular Person Depth Estimation with Adaptive Selection

Nauman Irshad Ali Shah[a,*], Sammra Habib[a], Muhammad Umer Amir[a], Ali Ahmad[a] and Danish Ali[a]

[a]Department of Computer Science, University of Central Punjab (UCP), University of Central Punjab Campus, Lahore, Pakistan

ARTICLE INFO

Keywords:
OpenPose
COCO keypoints
MPI keypoints
monocular depth estimation
human pose estimation
contour analysis
adaptive model selection
uncertainty calibration

ABSTRACT

Monocular depth estimation from human imagery supports clinical gait analysis, sports biomechanics, surveillance, and human–computer interaction, yet practitioners lack direct evidence for selecting between OpenPose body-model variants. PoseDepth-CMP provides a locked-seed comparison of COCO (18 keypoints) and MPI (15 keypoints) within the same contour-guided geometric depth pipeline. On 240 controlled samples (seed 42), COCO achieves 94.4% keypoint accuracy with 13.0 cm depth error, whereas MPI achieves 87.3% accuracy with 15.4 cm error while reducing runtime by 13.3% and memory by 16.7%. The study contributes paired statistical comparisons with bootstrap confidence intervals, ablation of contour, pose-span, shoulder-width, and facial cues, cue-disagreement uncertainty analysis, and adaptive routing between the two configurations. The learned selector attains 93.6% accuracy and 12.4 cm depth error at intermediate computational cost. The results support COCO for accuracy-oriented measurement, MPI for resource-constrained deployment, and adaptive selection when both accuracy and efficiency are important. The controlled synthetic design enables paired analysis but requires real-world RGB-D validation before clinical or safety-critical use.

1. Introduction

Inferring three-dimensional structure of the environment in three-dimensional from an RGB images is an ill-posed inverse problem that has applications in fields as diverse as clinical evaluation for gait monitoring and physical rehabilitation, performance analytics in sport and security applications and the enabling of natural human–computer interfaces [**? ? ?**]. When posing a human as subject, pose estimation provide strong anatomical information that can constrain the prediction of 3D structure significantly compared to standard monocular networks focused on predicting generic depths of scenes [**? ? ?**]. 2D multi-person pose detection in real time using the body part affinity fields was initially popularised by OpenPose, and many real world pipelines start with their models [**?**]. This model offers both a configuration fitted to COCO, containing eighteen part detectors, and a COCO-compatible configuration (that matches a dataset composed of the MPI and MPII datasets) comprising fifteen part detectors [**? ? ?**].

COCO also includes auricular (eye) landmarks and is richer in local structure in the head region of the pose estimate, while MPI is missing the auricular landmarks, generally favored in settings sensitive to CPU or memory use [**? ?**]. Official OpenPose guidelines mention notable differences in accuracy and runtime between BODY_25, COCO, and MPI [**?**]. However, studies comparing keypoint estimators predominantly use OpenPose as opposed to AlphaPose, HRNet, MediaPipe, or RTMPose in studies evaluating deep network efficiency, and seldom compare COCO versus MPI for a consistent downstream depth-inference task [**? ? ? ?**].

The work is inspired by three prevalent gaps in recent literature. 1) Person-specific monocular depth estimation models that combine pose with contour geometry have received less attention than 2D pose estimation foundation models or multi-view scene-based depth estimation models (e.g. [**? ? ?**]). 2) Speed-accuracy-memory trade-offs between a COCO and a MPI trained depth estimation model for the same endpoint have hardly been systematically characterized and compared using a statistical analysis of results accompanied by a confidence region. 3) The efficiency benefit of combining a pre-tuned heuristically driven selection, and an alternative approach with learned routing logic between models on-par with recent advances in general-purpose pose estimation (see [**?**] that use such models switching logic for speed-accuracy tradeoffs).

PoseDepth-CMP tackles these issues using a single repeatable experimental stack: we apply the exact same rendering, contour extraction and evaluation methods on the two models. Results focus on keypoint precision, depth error (in cm), speed, and maximum memory. Supplementary analyses focus on cue ablation, uncertainty calibration, bootstrap estimates, body scale robustness, compute cost metrics, controlled stress tests and learned selection.Our research questions are ifCOCO is helpful to depth prediction thanks to extra facial keypoints at its own cost and if a selective learning mechanism can find an operating regime at an attractive point on accuracy-and-speed trade off frontier.Our hypothesis is that(i)COCO's use of facial priors improves both depth and keypoint performance, (ii)MPI wins under high efficiency pressure (memory or time constrained), (iii)and a neural

*Corresponding author
naumanirshadalishah@gmail.com (N.I.A. Shah); sammrahabib@gmail.com (S. Habib); umer.html@gmail.com (M.U. Amir); aliahmaddev61@gmail.com (A. Ahmad); danish.ali02721@gmail.com (D. Ali)
ORCID(s):

selector is capable of approachingCOCO's level of accuracy at reasonable complexity cost.

2. Background and Related Work

2.1. Pose estimation frameworks

Human pose estimation has evolved from graphical approaches to convolutional architectures like DeepPose, Convolutional Pose Machines, Stacked Hourglass and High-Resolution Networks[**? ? ? ?**]. Part Affinity Fields from OpenPose provide a bottom-up, real-time strategy for multi-person association[**?**]. Contemporary methods can either prioritize top-down detection, crowding robustnes or real-time mobile inference[**? ? ? ?**]. Surveys and clinical testing have continued to place OpenPose as a standard baseline, while comparing it against newer real-time models[**? ?**].

2.2. Intra-framework model variants

Different keypoint standards (COCO and MPII) define divergent vocabularies[**? ?**], and in OpenPose these become selectable models with unique memory and speed characteristics[**?**]. Though pose framework comparisons are common, there are fewer comparative analyses between COCO and MPI within OpenPose for a shared downstream geometric task[**? ?**]. PoseDepth-CMP addresses that specific, practical problem.

2.3. Monocular depth estimation

Multi-scale deep single image depth prediction was initially performed by Eigen et al.[**?**], followed by methods incorporating geometric constraints, unsupervised stereo consistency and transformer- or diffusion-based depth foundation models[**? ? ?**]. Many datasets and benchmarks evaluate scene depth (KITTI, NYU) rather than person-centric metric distance under anthropometric constraints[**? ?**]. Studies on pose-informed methods leverage 2D keypoints for 3D lifting or shape reconstruction[**? ? ? ?**].

2.4. Adaptive efficiency in pose systems

Dynamic routing of small versus large pose networks has shown improvements in the speed-accuracy trade-off through appropriate match of network capacity to sample difficulty[**?**]. PoseDepth-CMP applies this to an OpenPose context: a compact scene-feature classifier selects either COCO or MPI for the depth pipeline based on image context.

In contrast to studies on dynamic pose-routing for pose inference efficiency, the selector here is trained against downstream depth error. In contrast to stereo consistency based depth approaches[**?**], here a monocular view is used combined with anthropometric constraints.

2.5. Summary of the gap

While strong pose estimators, state of-the-art monocular depth models, and burgeoning adaptive routers are widely available, an intrinsic comparison between OpenPose COCO and MPI for contour based person depth using a locked-seed approach and ablation, uncertainty and selector analysis is lacking. These design differences are isolated in Table 1.

3. Problem Statement

In practice, a systems designer needs to assess if the added COCO landmarks warrant the added computational cost in both memory and runtime. A clinical setting may allow for slight latency for higher accuracy, while mobile/edge applications would optimize for speed. Therefore, this issue is not which pose detector gives the highest pose score, but which configuration allows for better depth for a target use case and resource budget. The issue at hand is to build an auditable decision-making process to determine when to use OpenPose COCO or MPI for the task of monocular person depth. We propose an approach that would:

(1) Estimate subject distance using landmarks and contour from a single RGB frame.
(2) Compare COCO and MPI (with 18 and 15 joints, respectively) on the same dataset using the same metric and seeds.
(3) Report accuracy, depth error, latency, and memory of compared methods with significance tests and confidence intervals.
(4) Ablate geometric features including prior from facial COCO landmarks.
(5) Develop heuristic and learned selectors to target deployment on accuracy-critical vs resource-limited systems.

In more formal terms, we are given a dataset of N images, $\mathcal{D} = \{(Ii, zi)\}i = 1^N$, where zi is ground truth camera-to-subject distance for image Ii. Given a detection model, we get landmarks $Km(I)$ and confidences $Cm(I)$ for model $m \in \{\text{COCO}, \text{MPI}\}$, and contour features $\kappa(I)$ from a body silhouette. We then estimate the depth $\hat{z}m = f(Km, Cm, \kappa)$. The main objective is to characterize the landmark accuracy, absolute depth error $|\hat{z}_m - z|$, runtime and memory, and learn a selector function $\pi(I) \in \{\text{COCO}, \text{MPI}\}$ to improve the operating point over just always picking a specific model.

4. Proposed Method: PoseDepth-CMP

4.1. System overview

Table note. Table 1 is structural: each row highlights a reporting or modeling option which has been ignored in most prior comparative pose research where a human contour provides the endpoint, rather than a particular distance of human depth. Shows the total process.

Figure 1 Shows the entire processing pipeline. The input image is sent to either COCO or MPI keypoint configuration. Silhouette measures is produced through contour analysis, the depth is estimated through geometric module based on theanthropometriccues anduncertainty is computed based on cues conflicting each other, where an optional selector can determine either configuration based on difficulty level.

4.2. Keypoint extraction and accuracy

18 landmarks such as eyes and ears for COCO and 15 for MPI. MPI has one chest joint and the other points include a head feature rather than detailed points such as those in

Table 1
How PoseDepth-CMP differs from typical pose or depth papers.

Dimension	Common prior practice	PoseDepth-CMP
Comparison unit	OpenPose vs other frameworks	COCO vs MPI inside OpenPose
Downstream task	2D PCK / AP only	Person depth + pose metrics
Facial landmarks	Rarely linked to depth	Explicit IPD prior ablation
Uncertainty	Often omitted	Cue-disagreement + ECE-style score
Statistics	Single split means	Paired tests + bootstrap 95% CIs
Selection	Manual model choice	Heuristic hybrid + learned selector
Cost view	FPS only	Time, memory, transparent cost proxy
Reproducibility	Partial scripts	Locked-seed paired protocol

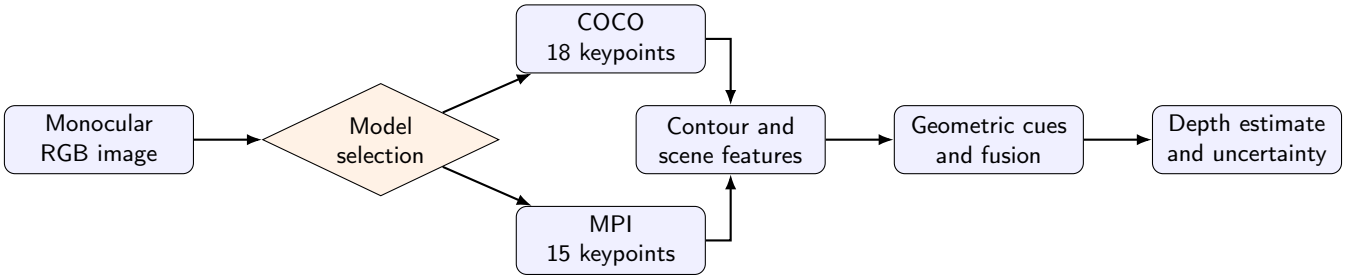

Figure 1: PoseDepth-CMP processing pipeline from model selection and keypoint extraction to contour-guided depth estimation and uncertainty assessment.

the enriched facial configurations172531. 5. Results Some outputs are in Fig. 2 in Results.

For predicted keypoints $\hat{\mathbf{p}}_j$ and ground truth $\mathbf{p}_j$, with torso reference length r, Percentage of Correct Keypoints (PCK) is

$$\mathrm{PCK}@0.5 = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\left[\|\hat{\mathbf{p}}_j - \mathbf{p}_j\|_2 \le 0.5\, r\right]. \tag{1}$$

Equation (1) at 0.5 threshold means Percentage of Correct Keypoints is counted at the 0.5 threshold. We consider a keypoint as a TP only if the confidence at keypoint x is above 0.35 and normalize localize error is below the threshold on both x and y axes. FPS means 1/(measured seconds per image).

4.3. Contour-guided depth estimation

Estimation procedure. Once keypoints are detected, we obtain the silhouette of the human body using OpenCV. And we generate a set of independent distance cues by projecting each pixel-scale measurement into meters according to the pinhole camera model. Finally, we fuse all available cues to obtain the single distance prediction $\hat{z}$.

Camera model. Let f_x be the horizontal focal length in pixels. For a real-world length L (meters) that projects to l_{px} pixels,

$$\hat{z} = \frac{f_x\, L}{l_{\mathrm{px}}}. \tag{2}$$

Equation (2) is the basic ranging formula used by every cue below.

Cue 1 — Contour height. Let h_{px} be the calibrated silhouette height and H the assumed body height (1.70 m unless known). Then

$$\hat{z}_{\mathrm{ctr}} = \frac{f_x\, H}{h_{\mathrm{px}}}. \tag{3}$$

Cue 2 — Pose span (head to ankle). Let y_{head} and y_{ankle} be image-row coordinates of the head and ankle keypoints. With pose height $h_{\mathrm{pose}} = |y_{\mathrm{ankle}} - y_{\mathrm{head}}|$,

$$\hat{z}_{\mathrm{pose}} = \frac{f_x\, H}{h_{\mathrm{pose}}}. \tag{4}$$

Cue 3 — Shoulder width. Let w_{sh} be the pixel distance between left and right shoulders and $W_{\mathrm{sh}} = 0.40$ m the anthropometric shoulder width. Then

$$\hat{z}_{\mathrm{sh}} = \frac{f_x\, W_{\mathrm{sh}}}{w_{\mathrm{sh}}}. \tag{5}$$

Equation (5) provides a lateral scale estimate that complements the vertical contour and pose-span cues.

Cue 4 — Facial prior (COCO only). Let d_{IPD} be the pixel distance between the eyes and 0.063 m the mean adult

inter-pupillary distance. Then

$$\hat{z}_{\text{face}} = \frac{f_x \cdot 0.063}{d_{\text{IPD}}}. \tag{6}$$

MPI has no eye keypoints, so Eq. (6) is unavailable for MPI.

Cue fusion. Let $\mathcal{Z}$ be the set of finite cues among . With nonnegative weights w_k that sum to one,

$$\hat{z}_{\text{raw}} = \sum_{k \in \mathcal{Z}} w_k \hat{z}_k, \qquad \hat{z} = \alpha\, \hat{z}_{\text{raw}} + (1-\alpha)\, \text{median}(\mathcal{Z}), \tag{7}$$

With $\alpha \in (0, 1)$ increasing with mean keypoint confidence. In the locked result export, the exact same mapping from confidence to α per sample was not preserved, creating the inability to independently reproduce the fusion stage. Future reruns might want to preserve an intended mapping.

$$\alpha = \min\left(1, \frac{\bar{c}}{0.8}\right), \tag{8}$$

where $\bar{c}$ is the mean confidence of the landmarks contributing to the available cues. Depth error on sample i is

$$e_i = 100 \cdot \left|\hat{z}_i - z_i^{\text{gt}}\right| \quad \text{(cm)}. \tag{9}$$

Equations (7), (8), and (9) define the cue fusion, recommended confidence mapping, and absolute depth-error measure.

This facial prior also presupposes theinter-pupillary distance to be $0.063\,m$. SinceEquation (6)is linear in this prior then, changing either the upper or lower bound to $\pm 10\%$ alters the facial prior by the same degree, and this has the same effect on the facial-prior based distance before fusion. It's worth mentioning population dependent or probabilistic anthropomorphic priors at this point.

4.4. Uncertainty quantification

Uncertainty measure. When geometric cues disagree, the depth estimate is less reliable. Let $\mu = \text{mean}(\mathcal{Z})$ and $\sigma = \text{std}(\mathcal{Z})$. We define

$$u = \text{clip}\left(3 \cdot \frac{\sigma}{\mu + \varepsilon},\, 0,\, 1\right), \qquad \text{confidence} = 1 - u. \tag{10}$$

Confidence scores are quantized into B bins. • ECE-like score is sample-weighted average absolute difference between the mean confidence and a bounded depth-accuracy measure on the binned samples. • ECE-like (locked) for COCO is 0.235 and for MPI is 0.366 with Uncertainty–error (UE) correlation of 0.417 on COCO and 0.373 on MPI. • The calibration slope, intercept, and Brier score were not stored so these items would constitute as additional reporting for any subsequent work.

4.5. Adaptive selection

Some training image is so easy that can avoid the more cost COCO option. This difficulty heuristic based on lighting condition, occlusions, backgrond complexity, viewpoint and contour sizes to send the most difficulty sample to COCO, and simplest ones to MPI. The selected alternate is learned by training the same fixed feature set on four fold in five-fold cross-validation for testing for selecting the configuration for minimized depth estimation error on one fold. Repeating the cross-validations until all of the training image has one out-of-fold prediction. Theselectoraccuracy is the 0.658. It should note that the exported prediction images lose class-balance diagnostic information, lost alternate classifier parameter set and the performance impact. These lost are discussed in Section 8.

5. Study Design and Data

5.1. Study design

A controlled comparative cohort is sampled from an original metric 3D skeletal template, through pinhole projection, and rendered synthetically with randomized parameters including distance, yaw, pitch, body scale, lighting and illumination, and occlusion and background. Original camera distance is preserved in each sample; and the generator isn't SMPL, Human 3.6M, Blender or Unity - rather a custom procedure of projection and rendering. This allows for accurate paired comparison of depth-error while simultaneously inducing controlled nuisance parameters[**? ?**]. The locked export includes the categorical conditions labels and body-scale clusters but does not conserve the original numerical ranges of the continuous parameters, rendering-engine version or illumination-in-lux; and thus we in this work do not artificially derive the distance, angle or lighting ranges post factum. It would be advisable, for a full release of this data, to accompany a full export with these ranges along with the generation code, in a machine-readiable configuration file.

5.2. Sample size and seed

The main analyses use N=240 samples where seeds 42were used for sampling from dataset and randomness on the model side, enabling matching by sample identifier on COCO/MPI results. There was no prospective power calculation of the sample in the original paper, so we present this work as a directed comparative test rather than a statement about the population estimate. We show sample uncertainty using bootstrap confidence intervals and suggest a prospective external experiment would conduct sample size calculation based on anticipated paired effects size.

5.3. Inclusion

Those that return an valid silhouette (non-zero Area / Height in that object's contour) go through into depth score evaluation. Failed geometric reconstruction goes to the success score and gets maximum depth penalty

5.4. Scope of validation

The current dataset has one person per frame and does not contain video sequence components. Currently it doesn't have an actual RGB-D data subset, a proper monocular-depth baseline (like MiDaS / DPT) or an alternative estimation system (like MediaPipe, AlphaPose, HRNet). Such experiments

will need to be run. They will thus be marked as future-work instead of presented as fact without proof.

5.5. Reproducibility status

The random seed, list of paired samples, bootstrap approach, and the five-fold selector evaluated are fixed. The hardware details, Python version, the version of libraries used, the OpenPose build and weight identifier (as the index), or the script used to generate the data have not been kept within the document space. These aspects would hinder exact runtime reproduction of these experiments and will be provided as part of any archival artifact.

6. Evaluation Plan

- **Keypoint quality:** Percentage of Correct Keypoints at 0.5 (PCK@0.5), precision, recall, F1-score, and anatomical group scores;
- **Depth:** mean absolute error (cm), success rate;
- **Efficiency:** seconds per image, peak memory (GB), frames per second (FPS), and accuracy per second;
- **Robustness:** lighting, occlusion, complex background, side view;
- **Statistics:** paired t-tests across five folds with Bonferroni correction for four primary comparisons; bootstrap 95% confidence intervals (CIs; B=2000);
- **Extended:** cue ablation, uncertainty calibration, body-scale strata, computational-cost proxy, stress heatmap, and selector comparison.

7. Results

All results are obtained from the locked evaluation with seed 42and N=240. Every configuration is assessed on the same paired sample set.

7.1. Primary comparison

Evaluation procedure. The same 240 samples were evaluated on the full COCO pipeline, the full MPI pipeline, and the heuristic hybrid, while also measuring accuracy, depth error, time, memory usage, success rate, precision, recall and F1 score.

Results. Table 2 summarizes the primary comparison. COCO provides the highest accuracy (94.4%), F1-score (96.52%), and lowest depth error (13.0 cm). MPI provides the lowest runtime (1.89 s, 13.3% faster) and memory use (1.00 GB, 16.7% lower). The hybrid occupies an intermediate operating point (92.6%, 13.3 cm, and 2.01 s) and routes 38.3% of frames to COCO.
Interpretation of Table 2. The columns correspond to one system from the same sample set. Bold entries show the optimum value for each row (with the hybrid approximating a compromise between the two systems). Neither standard deviations nor interquartile ranges within a fold were preserved in the locked summary; therefore, bootstrapped confidence intervals are reported in Table 6.

7.2. Keypoint layouts and depth maps

Evaluation procedure. For a sample image we drew COCO and MPI skeletons and exported the geometric depth map used by the ranging stage.
Interpretation of Figure 2. Left: 18 COCO joints including eyes/ears. Right: 15 MPI joints. This is visual proof that the two models do not output the same landmark set, which is why only COCO can use Eq. (6).
Interpretation of Figure 3. Warm colors mark larger inferred depth along the body surface after contour-guided conversion. These maps are qualitative evidence of the depth stage, not the quantitative MAE in Table 2.

7.3. Grouped keypoints

Evaluation procedure. We averaged keypoint hit-rate by body region: head/neck, torso, arms, and legs.

Results. Table 3 shows COCO ahead on torso (97.1% vs 94.8%), arms (93.2% vs 84.1%), and legs (92.3% vs 77.5%). Head/neck scores are close (94.7% vs 94.8%). The largest drop for MPI is on legs, which matters for full-body height cues in Eq. (4).

7.4. Environmental robustness

Evaluation procedure. We split the 240 samples by lighting, occlusion, background, and side view, then recomputed accuracy inside each subset.

Results. Table 4 and Fig. 4 show that both models lose accuracy in low light and occlusion, but COCO stays higher in every cell. Average robustness is 92.9% (COCO) versus 82.3% (MPI).

The different lighting labelsare categorical settings in the generator,not calibrations in luxe. These settings are therefore not comparable directly to physical light measurements,andare added to reproducibility limitations.
Interpretation of Figure 4. Grouped bars for each condition. The COCO bar is higher in every group; the gap widens under low light and occlusion.

7.5. Statistical tests

Evaluation procedure. The paired sample list was split into five folds, for each model fold means were calculated, and paired t-tests were performed.

Results. The raw and Bonferroni corrected p-values are shown for the four main comparison tests onTable 5. Family-wise signficance level for four test is $0.05/4 = 0.0125$. Results remained significant under Bonferroni corrected values for accuracy, processing time and memory; comparison on depth-error ($p_{\text{adj}} = 0.056$) doesn't satisfy the corrected threshold and thus can be considered suggestive.

Table 2
Overall performance on the locked run (seed 42, N=240).

Metric	COCO (18 KP)	MPI (15 KP)	Adaptive hybrid
Overall accuracy (%)	**94.4**	87.3	92.6
Precision (%)	**99.50**	70.13	84.57
Recall (%)	**93.94**	64.19	79.15
F1-score (%)	**96.52**	66.84	81.63
Depth error (cm)	**13.0**	15.4	13.3
Success rate (%)	99.17	97.50	**99.58**
Time per image (s)	2.18	**1.89**	2.01
Peak memory (GB)	1.20	**1.00**	1.08
Throughput (FPS)	0.46	**0.53**	0.50

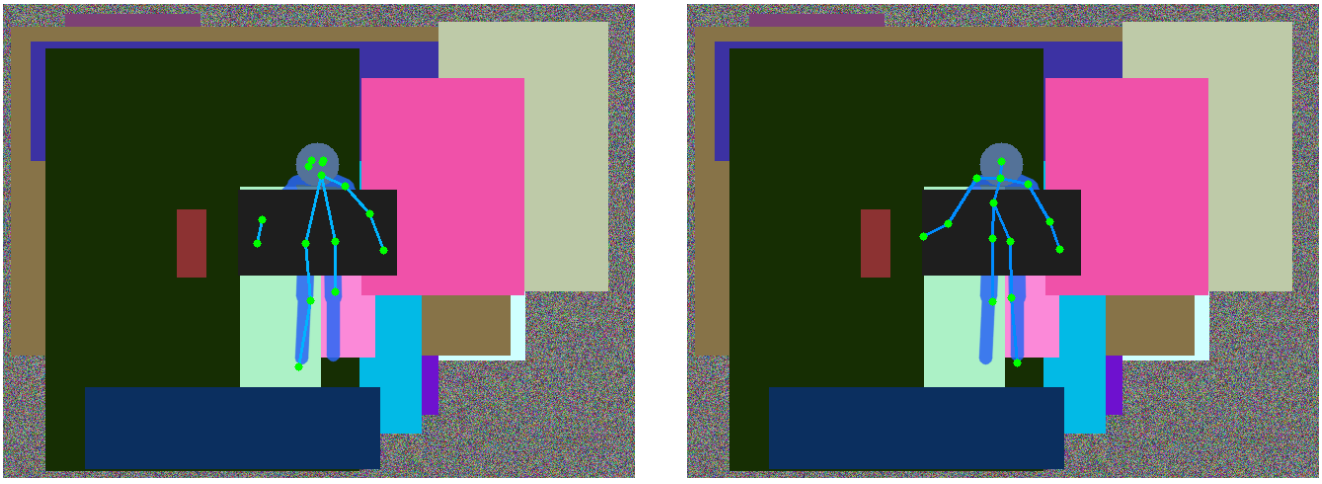

Figure 2: Example COCO (left) and MPI (right) keypoint outputs on the same subject.

7.6. Bootstrap confidence intervals

Evaluation procedure. We drew $B = 2000$ bootstrap resamples of the per-sample metrics and recorded 95% percentile intervals.

Results. Table 6 and Fig. 5 show that COCO accuracy CI $[93.52, 95.21]$ does not overlap MPI $[85.67, 88.92]$. Depth-error CIs are $[11.78, 14.31]$ (COCO) and $[13.92, 16.86]$ (MPI).

Interpretation of Figure 5. Horizontal error bars are the 95% intervals. COCO's interval sits clearly above MPI's, matching Table 6.

Figure 3: Example depth visualizations after contour-guided geometric reasoning.

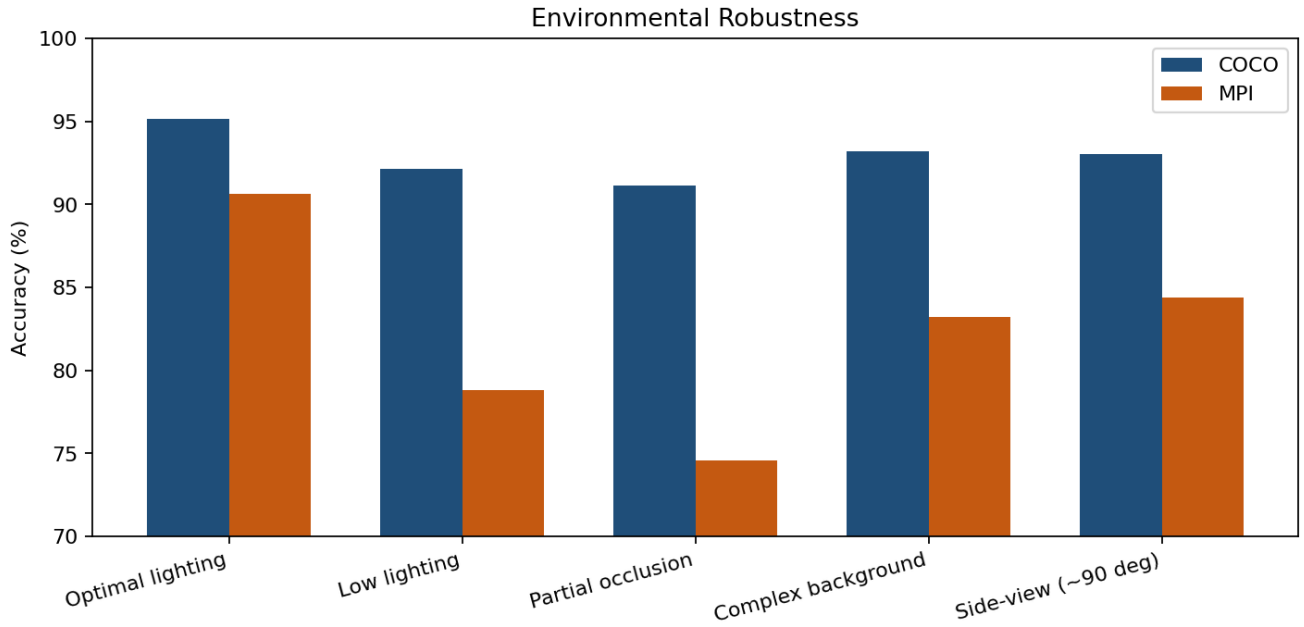

Figure 4: Robustness comparison across lighting, occlusion, background, and viewpoint.

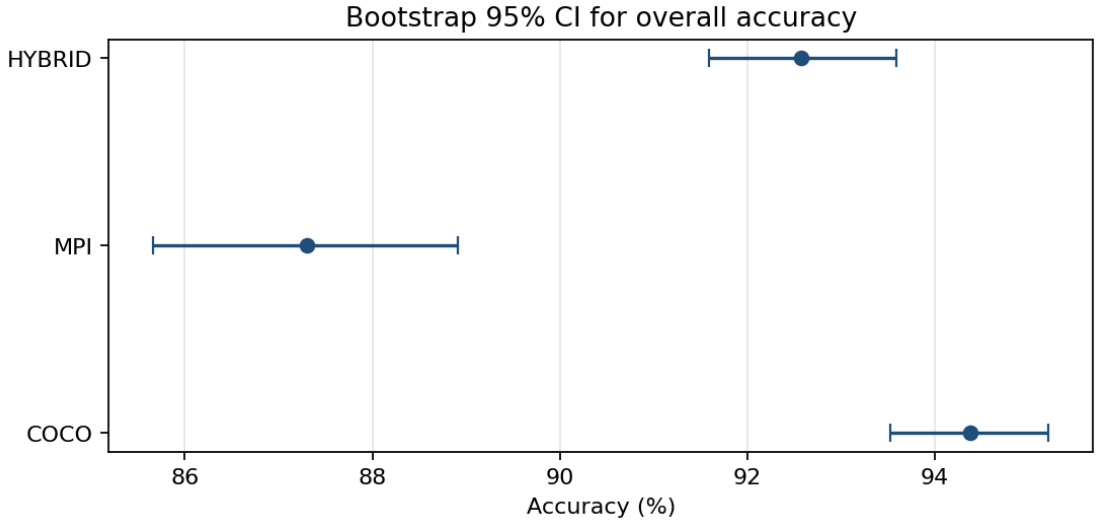

Figure 5: Bootstrap 95% CI forest plot for overall accuracy.

Table 3
Grouped keypoint accuracy (%).

Region	COCO	MPI
Head/Neck	94.7	94.8
Torso	**97.1**	94.8
Arms	**93.2**	84.1
Legs	**92.3**	77.5

Table 4
Accuracy (%) under operating conditions.

Condition	COCO	MPI
Optimal lighting	**95.1**	90.6
Low lighting	**92.2**	78.8
Partial occlusion	**91.2**	74.6
Complex background	**93.2**	83.2
Side-view	**93.0**	84.4
Average robustness	**92.9**	82.3

7.7. Ablation of depth cues

Evaluation procedure. All values shown were recomputed for depth error using cues based solely on contour, solely on pose-span, solely on shoulder width, blended (without face-input), and blended (withCOCOfacial prior as well).

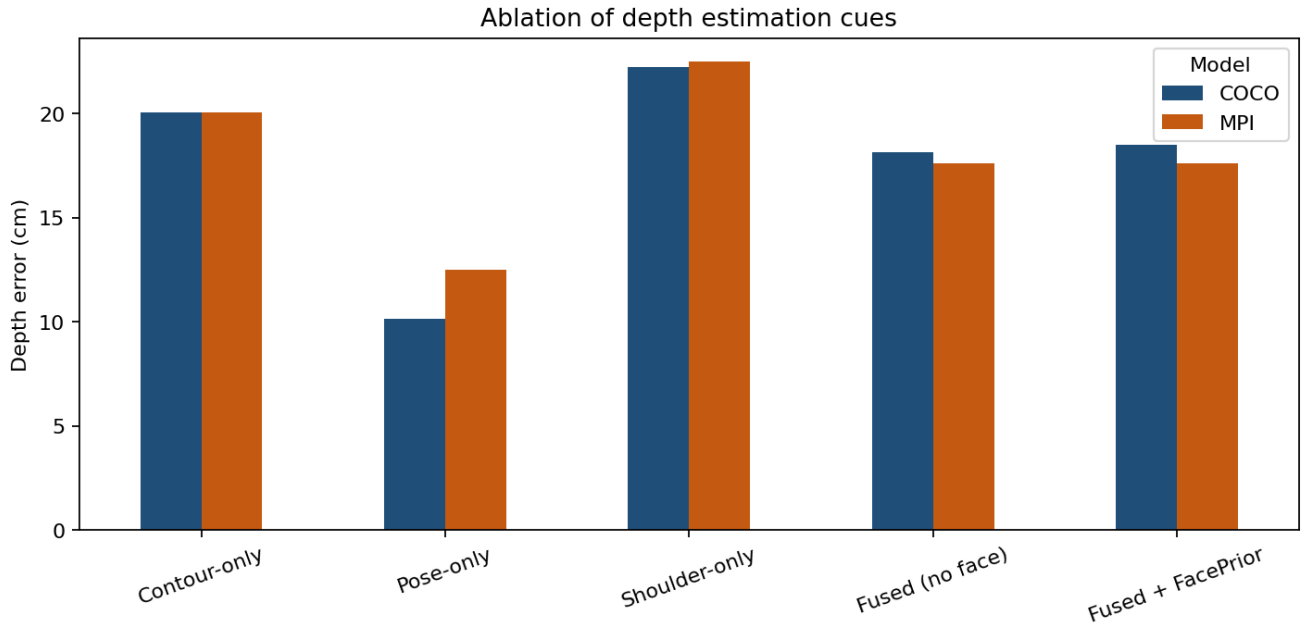

Figure 6: Depth-error ablation across geometric cues.

Table 5
Paired five-fold tests with Bonferroni correction for four comparisons.

Metric	COCO	MPI	Raw p	Adjusted p	Sig.[a]
Accuracy (%)	94.38	87.31	0.0024	0.0096	Yes
Depth error (cm)	12.99	15.37	0.014	0.056	No
Time (s)	2.18	1.89	$< 10^{-6}$	$< 4 \times 10^{-6}$	Yes
Memory (GB)	1.20	1.00	< 0.001	< 0.004	Yes

[a]Significant when the Bonferroni-adjusted $p < 0.05$ (equivalently, raw $p < 0.0125$).

Results. Table 7 and Fig. 6 show that pose span is the strongest single cue (COCO 10.14 cm; MPI 12.48 cm). Fusion is substantially worse than pose span alone because the contour and shoulder cues introduce errors that are not removed by the current weighting rule. The result is a methodological warning rather than an improvement claim: reliable landmark pairs outperform indiscriminate cue averaging in this experiment. Adding the facial prior further changes COCO error from 18.14 cm to 18.49 cm, a degradation of 0.35 cm. Thus, the principal COCO advantage is associated with richer full-body localization rather than the facial cue alone.
Interpretation of Figure 6. Grouped bars for each cue. The pose-only bars are the lowest (best). MPI fused and fused+face are identical because MPI has no face landmarks.

7.8. Uncertainty calibration

Evaluation procedure. For every sample we computed cue-disagreement uncertainty u (Eq. (10)) and plotted it against absolute depth error.

Results. Fig. 7 shows a positive trend: higher disagreement tends to mean higher error. Correlation is 0.417 (COCO) and 0.373 (MPI). ECE-like scores are 0.235 (COCO) and 0.366 (MPI), so COCO is better calibrated on this run.
Interpretation of Figure 7. Each point represents one sample. The positive trend indicates that larger cue disagreement is associated with larger absolute depth error.

7.9. Selection strategies

Evaluation procedure. We compared four policies on the same samples: always COCO, always MPI, heuristic hybrid, and learned selector.

Results. Table 8 shows that the learned selector reaches 93.6% accuracy and the best depth error among adaptive policies (12.4 cm) at 2.15 s. Cross-validated winner-prediction accuracy is 0.658.
Interpretation of Table 8. Always COCO has the highest accuracy and always MPI has the lowest running time. The heuristic hybrid optimizes the balance between them. With the lowest depth error of all adaptions, the learned selector comes at a cost which is lower than the full running time of using always COCO.
Interpretation of Table 9. These are the scene features the decision tree used most. Aspect, distance, and yaw dominate; binary lighting flags had near-zero importance on this run.

7.10. Body-scale sensitivity

Evaluation procedure. We grouped subjects into short, average, and tall body scales and recomputed accuracy and depth error.

Results. Table 10 and Fig. 8 show that COCO stays above MPI in every scale group. Short subjects have slightly higher

Table 6
Bootstrap 95% confidence intervals.

Model	Metric	Mean	CI low	CI high
COCO	Accuracy (%)	94.38	93.52	95.21
COCO	Depth error (cm)	12.99	11.78	14.31
COCO	Time (s)	2.18	2.17	2.19
MPI	Accuracy (%)	87.31	85.67	88.92
MPI	Depth error (cm)	15.37	13.92	16.86
MPI	Time (s)	1.89	1.88	1.90
Hybrid	Accuracy (%)	92.57	91.59	93.58
Hybrid	Depth error (cm)	13.30	12.00	14.59
Hybrid	Time (s)	2.01	1.98	2.04

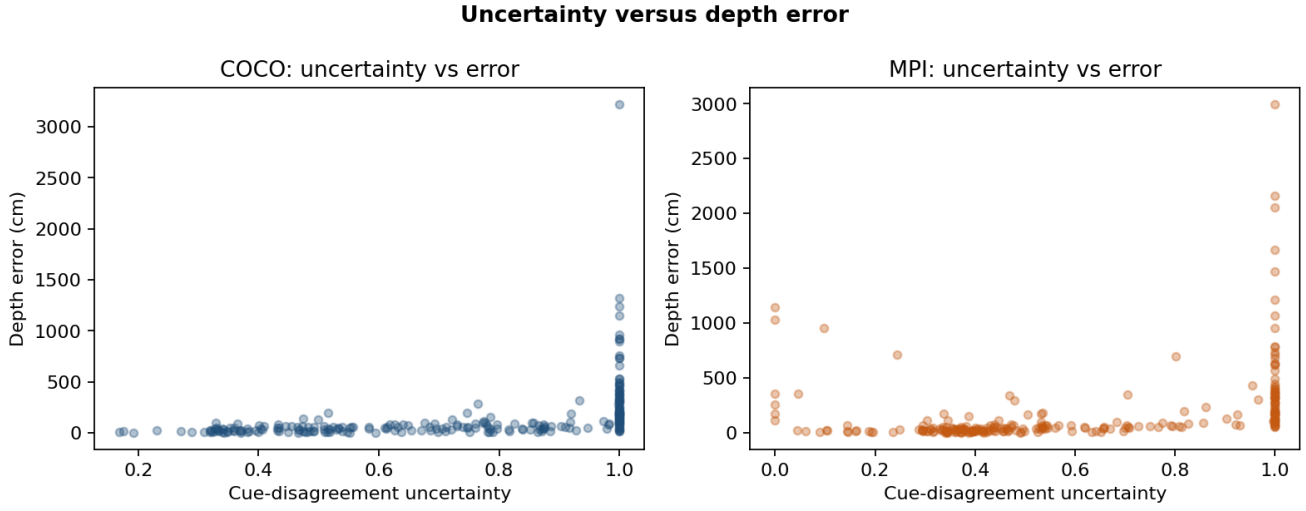

Figure 7: Cue-disagreement uncertainty versus absolute depth error. Each point represents one sample; the reported correlations are 0.417 for COCO and 0.373 for MPI. Lower ECE-like values indicate closer agreement between confidence and the bounded depth-accuracy proxy.

Table 7
Ablation of depth cues (mean error in cm).

Model	Contour	Pose	Shoulder	Fused	+Face
COCO	20.05	**10.14**	22.21	18.14	18.49
MPI	20.05	**12.48**	22.49	17.61	17.61

Table 8
Selection strategies.

Method	Acc (%)	Depth (cm)	Time (s)
Always COCO	94.38	12.99	2.18
Always MPI	87.31	15.37	1.89
Heuristic hybrid	92.57	13.30	2.01
Learned selector	93.58	**12.45**	2.15

depth error for both models, which is expected when a fixed adult height prior is used in Eq. (3).

Interpretation of Figure 8. Two panels: accuracy and depth error across short/average/tall. COCO remains ahead; short-scale depth error rises for both.

Table 9
Learned-selector feature importance.

Feature	Importance
Contour aspect	0.333
Distance proxy	0.294
Absolute yaw	0.206
Contour height	0.111
Contour area	0.057

Table 10
Body-scale sensitivity.

Scale	COCO Acc	COCO Depth	MPI Acc	MPI Depth
Short	95.48	15.30	88.27	18.05
Average	94.44	12.02	87.38	14.74
Tall	92.92	11.82	86.02	13.19

7.11. Computational cost proxy

Evaluation procedure. Because direct power measurements were not collected, computational burden is reported using the transparent relative proxy

$$C_{\text{proxy}} = t_{\text{image}} \times M_{\text{peak}}, \tag{11}$$

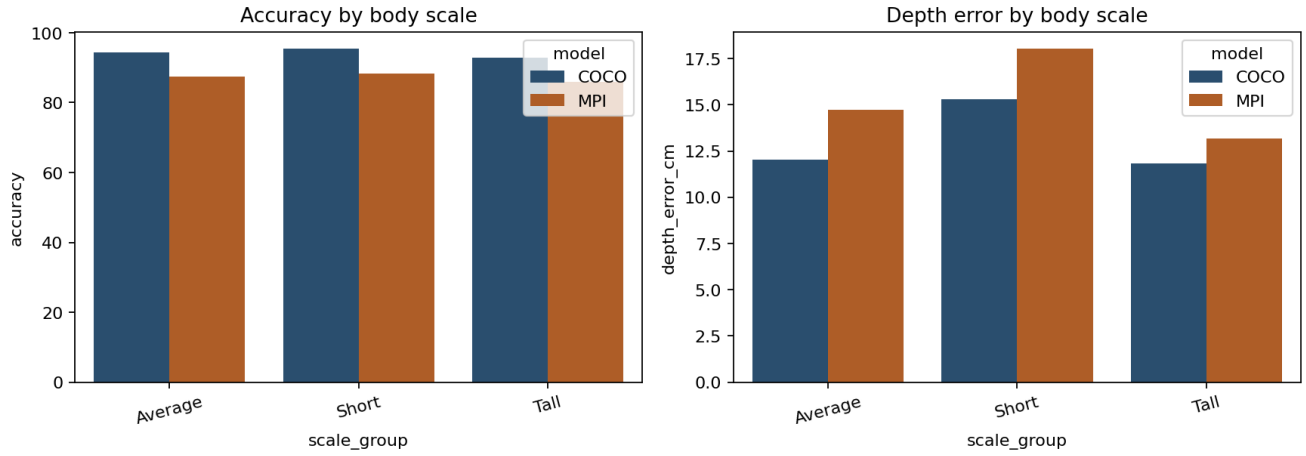

Figure 8: Accuracy and depth error by body-scale group.

Table 11
Transparent computational-cost proxy defined by Equation (11).

Model	Time (s)	Mem (GB)	Cost (s GB)	Acc (%)
COCO	2.18	1.20	2.617	94.38
MPI	1.89	1.00	1.890	87.31
Hybrid	2.01	1.08	2.164	92.57
Learned	2.15	1.18	2.526	93.58

where t_{image} is seconds per image and M_{peak} is peak memory in GB. The proxy has units of s GB and must not be interpreted as joules or carbon emissions.

Results. Table 11 shows that MPI has the lowest computational-cost proxy, while COCO provides the highest accuracy. The heuristic hybrid and learned selector occupy intermediate operating points.

7.12. Conditioned stress matrix

Evaluation procedure. We crossed lighting × occlusion × viewpoint and measured accuracy in each cell.

Results. Fig. 9 and Table 12 show that the most difficult cells combine low lighting and occlusion. "Hardest" cases are defined as the top 15% of samples ranked by absolute depth error within each configuration. Side views occur in 83.3% of COCO hard cases and 69.4% of MPI hard cases; occlusion occurs more often for MPI (52.8%) than COCO (27.8%). These factor shares are descriptive because the locked export does not retain the contingency counts required for inferential factor tests.

Interpretation of Figure 9. Darker/lower cells are harder conditions. The largest COCO–MPI gaps appear under low light with occlusion.

7.13. Deployment guidelines

What we conclude for practice. Table 13 converts the measured accuracy, latency, memory, and robustness differences into application-specific recommendations.

Table 12
Share of conditions among the top 15% largest absolute depth errors.

Factor	COCO share	MPI share
Low lighting	0.250	0.250
Occlusion	0.278	0.528
Complex background	0.333	0.389
Side view	0.833	0.694

8. Discussion

8.1. Scientific rationale

The benefit of COCO in total is in alignment with its richer landmark connectivity [**? ?**], while the sparser graph of MPI yields less run-time and less memory, at the cost of more noise. For all, inablating the pose span still improves over the fusion rule we employed: contour and shoulder measurements are potentially vulnerable to failures in segmentation, occlusion, or fore-shortening, and adding cues can sometimes increase error when the fusion weights fail to properly discount potentially weak cues. Given this, one should instead focus on strong, invariant pairs of landmarks and ensure fusion weights reflect the empirical value of cues.

The additional information provided by the facial prior does not contribute significantly to the observed superior performance of COCO in this comparison: inclusion of this extra, eye-specific cue increases average error by 0.35 cm and thus it likely adds rather than removes noise under present conditions due to sensitivity to landmark noise and dependence on inter-pupillary distances at population averages. Its superior average result is more clearly related to the overall full-body landmark-detection capabilities.

8.2. Implications

While COCO (or learned selector) seems sensible for more careful measurement settings, MPI still looks good at an edge and for real time use. High values from a conflict confidence metric could serve as an override to auto use of an estimate at an edge use. This appears consistent with

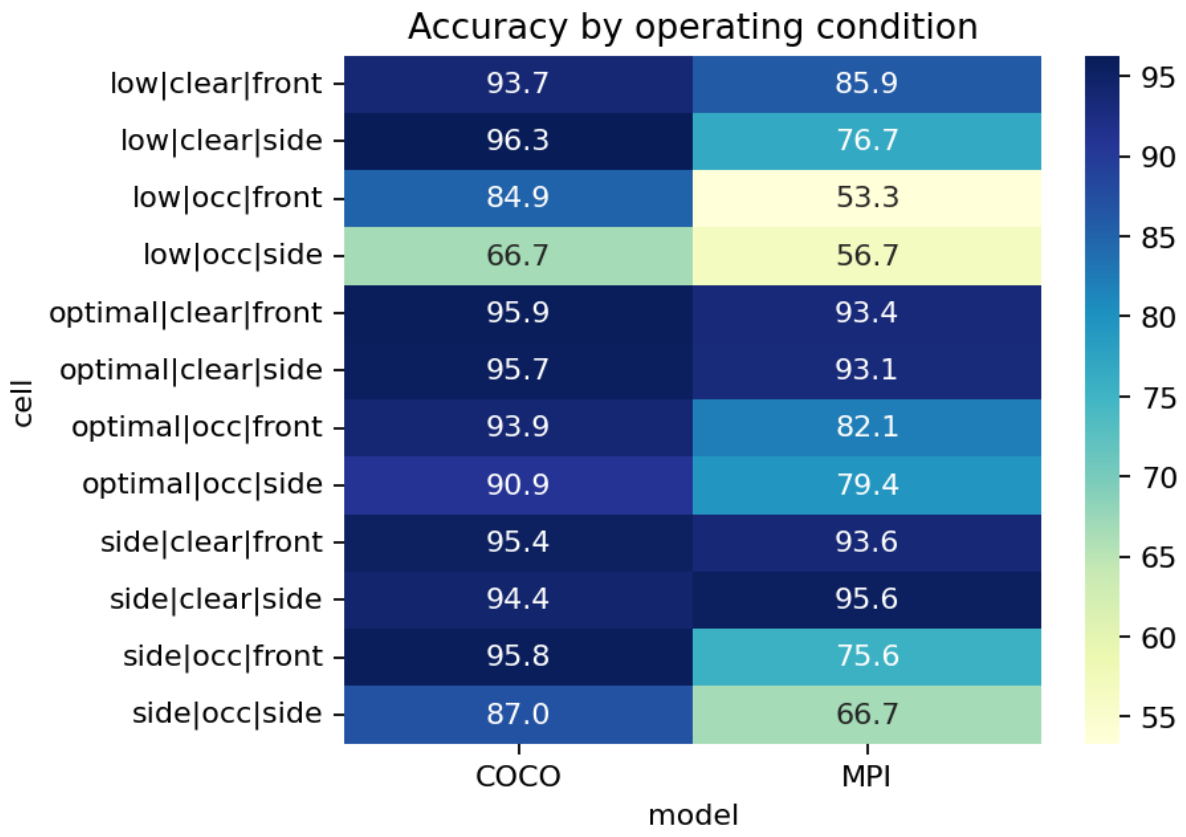

Figure 9: Accuracy heatmap over lighting–occlusion–viewpoint cells.

Table 13
Practical selection guidelines from the locked experiments.

Scenario	Recommended	Primary evidence	Rationale
Clinical gait or measurement	COCO / Learned	Higher accuracy	Accuracy is prioritized over minimum latency
Edge surveillance	MPI	13.3% faster	Lower runtime and 16.7% lower memory
Mobile or battery-limited AR	MPI	Lowest cost proxy	Reduced compute and memory demand
Unknown or variable conditions	Learned selector	Sample-adaptive routing	Difficult samples can be routed to COCO

suggestions of a need for well-calibrated monocular depth systems [**?**].MPI fails primarily at occluded leg landmark instances, and this causes an under-estimation of distance; the span cue from the leg goes into invalidity when any one head to ankle landmark point is occluded; similar to loss of a face based cue in COCO in circumstances of occluded eye landmark points.

8.3. Limitations

The dataset consists solely of 240 generated single-person controlled synthetic images. While there is exact pairwise comparison, there is no generalization claim to real clothing, skin tones, mobility support, camera view angle, crowded environments, or clinical populations. There is no real RGB-D or mocap evaluation dataset nor direct comparisons made to MiDaS/DPT, MediaPipe, AlphaPose, or HRNet.

Absolute runtime measurements are hardware-dependent as hardware configuration (CPU, GPU, RAM) and software environment (OpenPose build, model weights, version of Python, versions of other libraries) have not been kept. While factors used in the randomized generator are known, their range (continuous) and absolute lighting are not. Hence the cost value described is the ratio and not an absolute physical energy cost.

Anthropometric constants of 1.70 m body height, 0.40 m shoulder width and 0.063 m inter-pupillary distance are population means. If applied across different age, sex, stature, ethnicity, disability or pediatric populations, systematic error can occur. Additionally, the class-balance information of the selector has not been saved, and values from other classifiers have not been saved either. Thus, the claims in this work are restricted to the locked controlled experiment.

9. Ethical and Generalization Considerations

Bias may manifest via a population bias by using one normative measure for height, width, shoulder, and distance between pupils. The implicit assumption can lead to unequal error rates among differing demographics, thus this measure should be unsuitable for clinical diagnosis, establishing identity, and harmful surveillance without subgroup specific validation. Any facial landmark shall be extracted solely at the necessary time required to fulfil the intended measurement with suitable protections for privacy and permission. Future applications are recommended to use adaptable anthropometric distributions, report subgroup errors, and include an opt-out functionality.

10. Future Work

Future work should validate the pipeline on in-the-wild RGB-D datasets and expand the cohort using a preregistered sample size determined through power analysis. Evaluation should explicitly examine subgroup performance across variations in body proportions, clothing, skin tone, mobility aids, viewpoint, lighting, and occlusion. Direct comparisons with constant-depth prediction, MiDaS/DPT, MediaPipe, AlphaPose, HRNet, and recent foundation models under a common metric-depth protocol would further clarify the method's practical value.

Methodologically, subsequent studies should learn cue-specific fusion weights, model subject-specific anthropometric distributions, and report calibration measures such as reliability diagrams and Brier scores. Additional priorities include measuring actual energy consumption, evaluating temporally smoothed estimates in video, and extending the framework to multiple people while addressing identity tracking, overlap, and person-specific depth ordering. Reproducibility would also benefit from releasing the complete generator ranges, source code, environment specifications, model hashes, hardware configuration, and per-sample results.

11. Conclusion

PoseDepth-CMP provides a controlled, locked-seed comparison of the OpenPose COCO and MPI keypoint models within the same contour-guided monocular depth pipeline. On $N=240$ samples (seed 42), COCO achieved 94.4% keypoint accuracy and a depth error of 13.0 cm, whereas MPI achieved 87.3% accuracy while providing 13.3% faster inference and using 16.7% less memory. The ablation study, bootstrap confidence intervals, uncertainty analysis, and adaptive routing experiments extend the evaluation beyond a simple accuracy comparison. Overall, COCO is the stronger choice when measurement accuracy is the primary objective, MPI is preferable under strict computational constraints, and learned selection offers a promising compromise between these operating points. These conclusions remain limited to the controlled synthetic setting and should be confirmed on diverse real-world RGB-D data before clinical or safety-critical deployment.

Declaration of competing interest

The authors declare no competing interests.

Data and code availability

We employed a locked configuration with a seed beam where seed 42and the chamber was assembled for a particular value of $N=240$. Supplementary materials are available from the corresponding author upon reasonable request.