International Journal of Computer Assisted Radiology and Surgery (2026) 21:1217–1225
https://doi.org/10.1007/s11548-026-03610-6

ORIGINAL ARTICLE

A multi-view pipeline and benchmark dataset for 3D hand pose estimation in surgery

Valery Fischer[1,2] · Alan Magdaleno[1,2] · Anna-Katharina Calek[1] · Nicola Cavalcanti[1] · Nathan Hoffman[1] · Christoph Germann[1] · Joschua Wüthrich[1] · Max Krähenmann[1] · Mazda Farshad[1] · Philipp Fürnstahl[1] · Lilian Calvet[1]

Received: 15 January 2026 / Accepted: 19 March 2026 / Published online: 11 May 2026
© The Author(s) 2026, modified publication 2026

Abstract

Purpose Accurate 3D hand pose estimation supports surgical applications such as skill assessment and robot-assisted interventions. However, surgical environments pose severe challenges, including intense and localized lighting, frequent occlusions by instruments or staff, and uniform hand appearance due to gloves, combined with a scarcity of annotated datasets for reliable model training.

Method We propose a multi-view pipeline for 3D hand pose estimation in surgical contexts that requires no domain-specific fine-tuning and relies on off-the-shelf pretrained models. The pipeline combines person detection, whole-body pose estimation, and 2D hand keypoint prediction on tracked hand crops, followed by constrained 3D optimization. We also introduce a benchmark dataset with over 68,000 frames and 3,000 manually annotated 2D hand poses with triangulated 3D ground truth, recorded in a high-fidelity physical operating room replica.

Results Quantitative evaluation on the proposed dataset shows that the pipeline consistently outperforms state-of-the-art baselines for both two- and three-dimensional hand pose estimation. The method substantially improves joint localization accuracy, achieving up to a 31% reduction in 2D joint error and a 76% reduction in 3D joint position error. The approach remains robust across increasing levels of scene complexity, including motion and partial occlusions.

Conclusion The proposed multi-view pipeline demonstrates the potential of combining robust detection, tracking, and geometric optimization for three-dimensional hand pose estimation in surgical environments without domain-specific retraining. Together with the introduced dataset, this work provides a baseline framework and benchmarking resource to support future research on surgical motion analysis and objective skill assessment.

Keywords Hand Pose Estimation · Benchmark Dataset · Surgery · Multi-view

Introduction

Automatic 3D hand pose estimation in the surgical environment opens up new possibilities for tracking surgical activity, enabling detailed analysis of hand motions that can be used for skill assessment [3], and surgical training [4, 5]. At the same time, 3D hand tracking has the potential to enable natural, touch-free interaction with surgical robots, imaging systems, and augmented or virtual reality platforms, while helping to preserve the sterile field by reducing or eliminating physical contact with control panels or non-sterile surfaces [6–8].

Approaches to 3D hand pose estimation can be broadly categorized into monocular and multi-view methods. Monocular methods [9, 10] are mostly learning-based pipelines that train deep neural networks on large annotated datasets to directly regress 3D keypoints or model parameters from RGB images. While powerful in controlled environments, these methods often exhibit limited accuracy in unseen scenarios and struggle to generalize to out-of-distribution environments such as surgical scenes. This limitation can be mitigated by incorporating additional information from pose priors [11, 12].

✉ Valery Fischer
vfischer@ethz.ch

1 University Hospital Balgrist, University of Zurich, Zurich, Switzerland

2 ETH Zürich, Zurich, Switzerland

In contrast, multi-view methods [13, 14] leverage multiple synchronized viewpoints, making the problem well-posed and resolving the scale ambiguity inherent in monocular estimation. They also demonstrate improved robustness to occlusion, making them the preferred solution for surgical applications where both robustness and accuracy are paramount. Typically, these methods consist of two main stages: (i) detecting 2D joint locations (*keypoints*) in each view and (ii) reconstructing the 3D configuration by optimizing the 2D keypoints under geometric and, optionally, anatomical constraints [15], where 2D hand poses are generally extracted using data-driven methods [1, 16].

Existing methods for 2D keypoint extraction face significant challenges under surgical conditions: Surgical lighting is very bright, focused on a small region, and produces high-contrast glare, as illustrated in Fig. 1. Occlusions are frequent, caused by instruments or other staff, and gloves create uniform hand appearance with reduced texture cues. Poor performances of state-of-the-art 2D pose estimation methods [1, 16] can be mostly attributed to the lack of suitable training datasets. On one hand, synthetic datasets such as RHD [9] provide dense annotations but fail to capture real-world visual noise and occlusions, as hands are typically rendered fully visible. On the other hand, most real datasets such as FreiHAND [17] and HandCo [18] lack 3D ground truth in realistic, cluttered conditions, or are limited to simple, highly controlled acquisition setups, while [5] is restricted to egocentric views only. Similar domain shift issues occur for whole-body pose estimation models that include hand predictions [1, 16, 19]. Srivastav *et al*. [20] proposed adapting instance detection and pose estimation models to surgical settings but are restricted to body poses only.

In this work, we propose a robust, training-free pipeline for 3D hand pose estimation in surgical environments that overcomes domain shift and occlusion challenges without requiring retraining or fine-tuning. Our design is motivated by three key observations: (i) Modern object detectors such as YOLOv11 [21] can reliably detect humans even under unseen conditions. (ii) Recent tracking models, notably Efficient Track Anything [22], enable robust long-term tracking under occlusions, appearance changes, and temporary detection failures. Their object-agnostic design makes them more resilient to domain shifts than hand-specific models, a property we explicitly leverage in surgical scenarios. (iii) State-of-the-art 2D hand pose estimators achieve highly accurate joint localization when applied to well-cropped hand regions.

These insights motivate a top-down multi-stage pipeline, shown in Fig. 2, that extends the principle of top-down 2D body pose estimation, namely detecting persons first and then estimating poses inside their bounding boxes, to the robust detection and accurate localization of hand joints in surgical video without domain-specific retraining. The pipeline first detects individuals using YOLOv11, applies full-body 2D pose estimation within each bounding box, identifies keyframes with confident hand detections, and then tracks hands using Efficient Track Anything to maintain temporal consistency despite intermittent detections. Finally, hand keypoints are refined using a high-resolution 2D hand pose estimator.

From the resulting multi-view 2D detections, 3D hand poses are reconstructed through triangulation followed by constrained optimization. The optimization enforces reprojection consistency, temporal smoothness and shape consistency, producing robust and accurate hand trajectories.

In summary, our main contributions are threefold. First, we present a modular, training-free pipeline for accurate multi-view 3D hand pose estimation in surgical environments. Second, we introduce a novel real multi-view surgical dataset comprising over 68,000 frames and 3,000 annotated 2D hand poses with corresponding triangulated 3D ground truth. Third, we conduct a comprehensive evaluation, including quantitative and qualitative comparisons with state-of-the-art methods as well as detailed ablation studies.

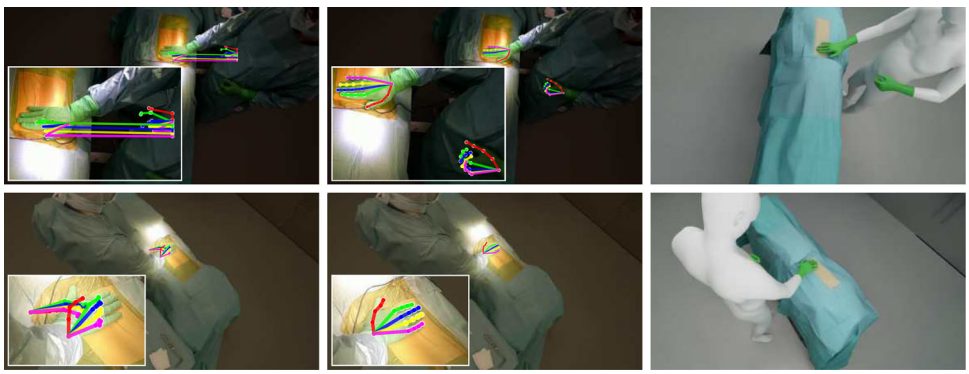


Fig. 1 Example showing the challenges of hand localization in surgical scenes caused by lighting conditions and the similar appearance of gloves and gowns. Left: RTMPose [1] fails to accurately localize the hands. Middle: our 2D predictions. Right: SMPL-X [2] 3D reconstruction using our 2D predictions from the same viewpoints

Method

Problem formulation

We consider the task of estimating accurate 3D hand poses in multi-view surgical videos captured by a set of calibrated and synchronized RGB cameras. Let the set of V synchronized cameras be denoted as $\mathcal{C} = \{C_1, \ldots, C_V\}$, each providing a video stream $\mathcal{V} = \{V_1, \ldots, V_V\}$ of resolution $W \times H$. The sequence contains N synchronized frames, with M_t unique hands visible at time t, each represented by J anatomical joints.

At each time step t, the 3D position of joint j for hand m in world coordinates is denoted by $\mathbf{X}^m_{t,j} \in \mathbb{R}^3$, and its 2D projection in camera view v by $\mathbf{x}^m_{t,j,v} \in \mathbb{R}^2$. The projection function $\Pi_v(\cdot)$ encodes the camera intrinsics and extrinsics, such that $\mathbf{x}^m_{t,j,v} = \Pi_v\left(\mathbf{X}^m_{t,j}\right)$. Predicted quantities are written as $\hat{\mathbf{X}}^m_{t,j}$ and $\hat{\mathbf{x}}^m_{t,j,v}$ for 3D and 2D joint locations, respectively. The problem can be decomposed into two subproblems.

The first is *2D hand pose estimation*, where given the set of multi-view RGB frames $\{V^v_t\}_{v=1}^V$, the goal is to detect and localize 2D hand joints for all visible hands. Each view v provides a set of detected hand bounding boxes $\mathcal{B}_{t,v}$, where $b^k_{t,v} = (x_{\min}, y_{\min}, x_{\max}, y_{\max})$ denotes the k-th detected region. Within each bounding box, a 2D hand pose estimator predicts a set of joint keypoints $\mathcal{P}^k_{t,v} = \left\{\left(\hat{\mathbf{x}}^k_{t,j,v}, c^k_{t,v,j}\right)\right\}_{j=1}^J$, where $c^k_{t,v,j} \in [0, 1]$ represent their confidence scores.

The second is *3D hand pose estimation*, where given the associated 2D joint detections $\hat{\mathbf{x}}^m_{t,j,v}$ and calibrated camera parameters, the goal is to estimate the corresponding 3D joint configuration $\mathcal{Q}^m_t = \left\{\hat{\mathbf{X}}^m_{t,j}\right\}_{j=1}^J$ for each hand m.

This formulation defines the 3D hand pose estimation task as the sequential resolution of 2D joint localization and 3D reconstruction under multi-view geometric consistency, which we adopt in this work. We describe our method for 2D pose estimation in Sect. 2D hand pose estimation and explore different sets of constraints from literature for the 3D pose estimation problem in Sect. 3D hand pose estimation.

2D hand pose estimation

Figure 2 provides an overview of our pipeline. In this section, we focus on the 2D pose estimation task (left part of the schematic), whose goal is to obtain accurate 2D hand keypoints in surgical scenes, where hands often appear small, are partially occluded, or undergo drastic illumination changes that obscure visual cues. To address these challenges, we adopt a modular, top-down strategy that combines reliable body detection, robust tracking, and high-resolution hand pose refinement. In the scope of this study, we focus on the case of a single surgeon in the scene. The overall process comprises three main components: (i) whole-body pose estimation for coarse initialization of hand locations, (ii) hand tracking from confident detections using Efficient Track Anything to maintain continuous and robust hand bounding boxes, and (iii) refined 2D hand pose estimation using a specialized model applied at near-native resolution for accurate localization.

Whole-body pose estimation and initialization From synchronized multi-view input videos $\mathcal{V}$, we first detect the surgeon in each frame using YOLOv11 [21], a state-of-the-art detector trained on large and diverse datasets. Within each person bounding box, we apply the RTMPose-XL whole-body 2D pose estimator [1] to predict full-body keypoints,

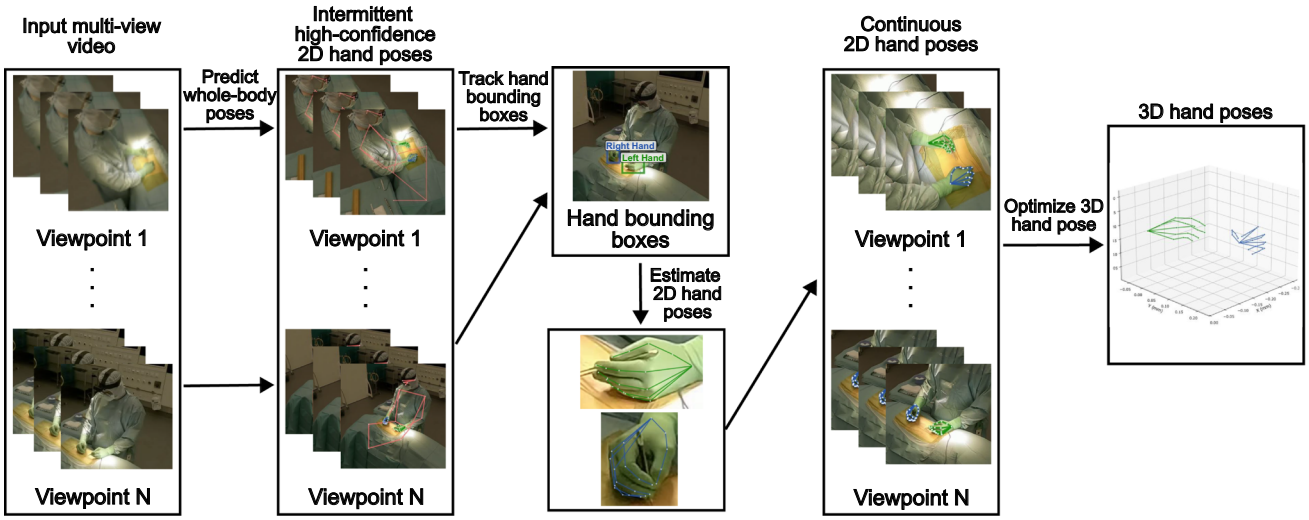


Fig. 2 Overview of our 3D hand pose estimation framework. Given multi-view videos from the OR, we first detect and track persons and then predict full-body and refined hand poses. Finally, we optimize a sequence of 3D hand poses consistent with anatomical and temporal constraints

including the hands. These detections provide coarse hand locations and bounding boxes, which serve as initialization for the dedicated hand-tracking stage. While these estimates may be imprecise due to the small image area of hands or strong lighting variations, they offer reliable priors for subsequent steps.

Hand tracking using Efficient Track Anything Keyframe selection. A set of keyframes in camera v is chosen based on the average confidence $c^m_{t,v}$ of all hand joint confidence scores from the whole-body model. For each hand, the frame with the highest $c^m_{t,v} > 0.3$ initializes tracking with Efficient Track Anything (1B-parameter TAM model), which is then propagated forward and backward. When tracking fails or reaches sequence boundaries, all successfully tracked frames for that hand are removed from the set, and the process restarts from the next highest-confidence frame until no valid detections remain. This propagation maintains consistent hand identities and continuous localization even during occlusions, rapid movements, or temporary detection failures. This process results in robust and temporally coherent hand trajectories $\mathcal{B}_{t,v}$ throughout the sequence. The Efficient Track module (based on SAM2) supports both point- and bounding box prompting followed by segmentation mask propagation. In our implementation, the bounding boxes defined by the predicted 2D hand keypoints from the whole-body pose estimator (as described in the previous paragraph) in selected keyframes are used as prompts to obtain reliable hand masks from well-visible frames. Output segmentations are then robustly propagated by the tracking module.

Accurate 2D hand pose estimation Each tracked hand region, cropped according to its propagated bounding box $\mathcal{B}_{t,v}$, is passed to a specialized 2D hand pose estimator for fine-grained keypoint prediction $\hat{\mathbf{x}}^m_{t,j,v}$. Operating under or near the native image resolution allows the model to recover detailed joint locations that are lost in the whole-body estimation stage. This refinement step effectively mitigates the low-resolution limitation of whole-body methods and provides precise, temporally stable 2D hand keypoints that serve as high-quality input for the subsequent multi-view 3D hand pose optimization. In order to mitigate propagation of suboptimal 2D hand poses into 3D, individual keypoints are filtered by the confidence scores produced by the hand model. Joint predictions associated with confidence scores $c^k_{t,v,j} < 0.1$ are discarded. In our implementation, we use the state-of-the-art 2D hand pose estimator RTMPose [1]. The exact model configuration and training setup are detailed in Section 4 of the Online Resource 1.

3D hand pose estimation

To obtain temporally consistent 3D hand trajectories, we formulate the 3D pose estimation problem as a constrained optimization problem solved with the L-BFGS-B algorithm. We incorporate constraints from prior work, some previously used in other contexts, including temporal smoothness, shape consistency to ensure stable hand geometry, and optionally biomechanical priors. Given initial 3D hand poses derived from a canonical hand attached to the triangulated body pose, we refine them by minimizing a weighted sum of loss terms that encode these constraints:

$$L_{\text{total}} = \lambda_{\text{reproj}} L_{\text{reproj}} + \lambda_{\text{smooth}} L_{\text{smooth}} + \lambda_{\text{shape}} L_{\text{shape}} + \lambda_{\text{biomech}} L_{\text{biomech}} \quad (1)$$

where each loss term is described below. Each frame is optimized sequentially, using the result of the previous frame.

For all subsequent equations, the current frame is indexed by i. Let joint index $j = 1$ correspond to the wrist. We define the centered 3D hand pose as $\hat{\mathbf{X}}^{\text{centered}}_i = \hat{\mathbf{X}}_i - \hat{\mathbf{X}}_{i,1}$ that is, the 3D hand pose with the wrist joint translated to the origin. For hand m and joint j, the corresponding centered keypoint is denoted as $\hat{\mathbf{X}}^{m,\text{centered}}_{i,j}$.

Keypoint reprojection loss (L_{reproj}) This term enforces geometric consistency between the projected 3D joints and the observed 2D detections, similar to [14], weighted by the confidence of the 2D predictions:

$$L_{\text{reproj}} = \sum_{m=1}^{M} \sum_{v=1}^{V} \sum_{j=1}^{J} c^m_{i,v,j} \left\| \Pi_v(\hat{\mathbf{X}}^m_{i,j}) - \hat{\mathbf{x}}^m_{i,j,v} \right\|_2^2. \quad (2)$$

Temporal smoothness loss (L_{smooth}) To encourage continuity over time, we penalize large changes in both joint positions and hand configuration between consecutive frames:

$$L_{\text{smooth}} = \sum_{m=1}^{M} \sum_{j=1}^{J} \left\| \hat{\mathbf{X}}^m_{i,j} - \hat{\mathbf{X}}^m_{i-1,j} \right\|_2^2. \quad (3)$$

Shape consistency loss (L_{shape}) This constraint ensures that the hand maintains a consistent shape across time by minimizing a rotation-invariant difference between consecutive frames, conceptually aligned with the method of Hewitt *et al*. [13] for body pose:

$$L_{\text{shape}} = \left\| \hat{\mathbf{X}}^{\text{aligned}} - \hat{\mathbf{X}}^{\text{centered}}_{i-1} \right\|_F^2, \text{ where} \quad (4)$$

$$\hat{\mathbf{X}}^{\text{aligned}} = \hat{\mathbf{X}}^{\text{centered}}_i \mathsf{R}^T, \quad \mathsf{R} = \mathsf{V}\mathsf{U}^T,$$
$$\mathsf{H} = \hat{\mathbf{X}}^{\text{centered},T}_i \hat{\mathbf{X}}^{\text{centered}}_{i-1}, \quad \mathsf{U}, \mathbf{\Sigma}, \mathsf{V}^T = \text{SVD}(\mathsf{H}).$$

This corresponds to the orthogonal Procrustes problem, which computes the rotation R that best aligns the two centered point sets in the least-squares sense.

Biomechanical loss (L_{biomech}) We also study the potential contribution of a biomechanical loss, although such terms are often found to be less effective. This loss encourages

biomechanical plausibility through a differentiable function that enforces joint-angle limits and anatomical feasibility, following Spurr *et al*. [15].

Benchmark dataset

To evaluate our method, we introduce a novel dataset of surgical hand motions captured with a calibrated multi-view setup in a replica of an operating room that faithfully reproduces real surgical conditions.

The dataset comprises two recording sessions. The first, **SPP** (spinal preparation), captures a single-surgeon procedure during the preparation phase using 12 synchronized static GoPro Hero Black 12 cameras (4K, 30 fps, fixed shutter). The second, **SPI** (spinal instrumentation), records a multi-surgeon cadaveric spinal fusion surgery using 11 GoPro Hero Black 12 cameras. All cameras were synchronized through a custom PCB controller ensuring sub-frame alignment [23]. Internal and external calibrations were performed using the OpenCV library and [24].

Complexity levels To systematically organize scene variability and support controlled benchmarking, we define a hierarchy of scene complexity levels, denoted as $\mathbf{L}_0$ through $\mathbf{L}_5$. Each level progressively increases the difficulty of hand pose estimation by introducing factors such as occlusion, motion, and the number of participants.

$\mathbf{L}_0$ – Single-person (static): A single person performs simple actions while remaining mostly stationary and facing the surgical table, interacting with the patient using small tools. All hand joints are fully visible in at least two views. This setting provides clean, unambiguous visual data ideal for baseline evaluation.

$\mathbf{L}_1$ – Dynamic single-person (slow, small tools): A single surgeon performs movements with slow hand motion and uses small surgical tools. Moderate occlusions may occur, but the structure of the hands remains largely visible. This level introduces natural variability while preserving controlled conditions.

$\mathbf{L}_2$ – Dynamic single-person (fast, large tools): Faster hand and arm motion combined with the use of larger tools causes stronger occlusions. These factors challenge temporal consistency and 3D reconstruction accuracy.

$\mathbf{L}_3$ – Static multi-person: Several participants (up to three) remain within the field of view, with limited movement and minimal occlusions. The setting increases visual clutter but retains good observability across cameras.

$\mathbf{L}_4$ – Partial observability: Multiple participants interact dynamically, often causing hands or tools to be partially hidden. Some camera views lose full hand visibility, introducing realistic visual ambiguity and limited multi-view coverage.

$\mathbf{L}_5$ – Dynamic multi-person: The number of visible participants changes over time with frequent occlusions, entering and leaving the scene. This represents the most unconstrained and complex condition.

In this work, we focus on $\mathbf{L}_0$–$\mathbf{L}_2$, which capture increasing complexity. The higher complexity levels ($\mathbf{L}_3$–$\mathbf{L}_5$) are included for future benchmarking. $\mathbf{L}_0$–$\mathbf{L}_2$ complexity levels already represent challenging surgical conditions for which no reliable 3D hand pose estimation methods currently exist in the literature.

The dataset comprises 20 sequences (300 frames each) spanning complexity levels $\mathbf{L}_0$–$\mathbf{L}_5$, totaling over 68,000 synchronized multi-view frames.

Annotation details For each sequence, we manually labeled 10 keyframes per camera, uniformly distributed across the 300-frame sequence (3840 × 2160 pixels), using a custom annotation tool. This process resulted in approximately 3,000 hand pose annotations. The 2D joint labels were triangulated into 3D ground truth using camera parameters, and their reprojections are used as the ground-truth 2D hand poses for evaluation. Annotation quality was further verified and evaluated through multi-annotator cross-review to ensure consistency (Section 2.4, Online Resource 1). Across 558 instances labeled by four annotators, mean 2D and 3D variability remained below 4.7 pixels and 4.6 mm, respectively, demonstrating high inter-annotator consistency.

Experiments

We conduct a comprehensive evaluation of the proposed method on the 2D hand pose estimation task in isolation (using 2D and 3D metrics) and compare it against 2D hand pose baselines, evaluated either independently or in combination. We also analyze the impact of the studied loss terms on 3D optimization through an extensive ablation study, all performed on the proposed dataset across complexity levels $\mathbf{L}_0$ to $\mathbf{L}_2$.

Baselines Our 2D pose estimation evaluation includes both whole-body and hand-specific baselines. We consider Sapiens [16], RTMPose [1] (noted RTM), and DWPose [19] (noted DW) as whole-body baselines. In addition, we evaluate a hand pose model from RTMPose [1] (noted HM)—as used in our pipeline—applied to cropped images extracted from bounding boxes around the hand regions predicted by the whole-body models. This setup allows us to isolate and assess the contribution of the hand-tracking component in our approach. The notation $A \rightarrow B$ indicates that the whole-body method A is used to obtain the hand bounding boxes, and the hand pose estimation model B is then applied within those regions. We deliberately restrict our baselines to training-free approaches, as our goal is to benchmark methods that can be directly transferred to new surgical environments without fine-tuning. Moreover, due to strong intra-procedural correlations in our dataset, super-

vised training could lead to data leakage and artificially inflated performance.

Experimental scores We evaluate our method using standard metrics that jointly assess 2D and 3D accuracy. The mean per joint position error (MPJPE) measures the average Euclidean distance between predicted and ground-truth 3D joint positions, while the mean reprojection error (MRE) quantifies how well the predicted 3D joints project back into the 2D image space. The percentage of correct keypoints in 2D ($PCK_{2D}@\tau$) and in 3D ($PCK_{3D}@\delta$) represent the fraction of keypoints whose reprojection or Euclidean error is below pixel or millimeter thresholds τ and δ, respectively. We also report the mean PCK ($mPCK$), computed as the average of PCK scores across thresholds (5, 10, 20, 30 for 2D and 5, 10, 25, 50 for 3D), expressed in pixels or millimeters. For 2D detection performance, we compute the average precision (AP) following the COCO evaluation protocol [25]. Finally, the mean joint error (MJE) measures the average 2D Euclidean error between predicted and ground-truth joints. Together, these metrics provide a comprehensive evaluation of geometric accuracy, reprojection consistency, and keypoint localization in both 2D and 3D spaces.

2D pose results Quantitative results are presented in Table 1, alongside our baseline comparisons. Figure 3 illustrates representative qualitative examples comparing the baselines with our method. The results show that our method significantly outperforms state-of-the-art models in predicting 2D hand poses across all reported metrics. This demonstrates that our top-down approach, combined with the tracking component of the model (noted T), enables more accurate and stable 2D hand pose predictions. The improvement can largely be attributed to the tracker's object-agnostic design, which allows it to localize hands effectively without prior knowledge of the target object. This leads to more reliable bounding box predictions and, consequently, improved 2D hand pose estimation. These results are further supported by the 3D evaluations presented in Table 2, where our method shows a clear advantage over predictions obtained from whole-body estimators alone (first row) or when combined with hand-specific models (second row).

3D pose results Results are summarized in Table 3, with qualitative examples shown in Fig. 4. Each variant of our optimization method is evaluated independently to analyze the impact of individual loss terms through an ablation study. As expected, the temporal smoothness constraint consistently yields significant improvements. The shape consistency term provides comparable performance on average but proves particularly effective for fast hand motions in $\mathbf{L}_2$ sequences, where it mitigates the degradation of 2D keypoints. Therefore, it is retained in our baseline configuration. In contrast, including the biomechanical constraint decreases performance, likely because the most plausible biomechanical shape does not precisely match the multi-view 2D predictions, leading to lower evaluation scores.

Discussion

Limitations and future work Although the proposed multi-view pipeline demonstrates strong performance for 3D hand pose estimation in surgical environments, several challenges remain unaddressed. The current implementation focuses primarily on controlled single-person scenarios, while more complex settings involving multiple participants, frequent self-occlusions, and low hand visibility pose additional difficulties that our method does not yet handle effectively. Execution time is another limitation, as the optimization and tracking stages remain computationally expensive, requiring around 30–45 min for a 300-frame sequence on an RTX 2060 GPU. Future work will target faster, learning-based feedforward models trained using 3D annotations produced by our robust pipeline, paving the way to near real-time performance. Extending the system to multi-person and occlusion-heavy scenes also represent key research directions.

Clinical implications Reliable 3D hand pose estimation could play an important role in surgical training by enabling objective, quantitative assessment of technical skills. By analyzing hand trajectories, motion smoothness, and bimanual coordination, such systems may provide automated feedback to trainees, complementing subjective expert evaluation. This could enable longitudinal skill tracking, allowing performance to be monitored across multiple training sessions to objectively measure progress, and evaluate training effectiveness.

Conclusion We presented a multi-view pipeline for 3D hand pose estimation in surgical settings that operates without domain-specific training. The integration of a top-down approach coupled with robust tracking and pose refinement yields superior accuracy compared to state-of-the-art methods, even under challenging occlusions. Despite current runtime limitations, this work establishes a solid foundation for future real-time, learning-based systems for surgical motion understanding.

Online resources

The online version of this article contains supplementary material. Online Resource 1 provides additional technical details, extended experimental results, and further information on the proposed dataset. Online Resource 2 is a video presenting an overview of the full pipeline and qualitative results for the SPI and SPP experiments.

Table 1 Quantitative evaluation on the 2D pose predictions, comparing different state-of-the-art methods against our method. HM refers to the RTM Hand Model used during pose refinement. * indicates the 2D pipeline without 3D optimization. **Best scores**, Second Best Scores

Metric	Sapiens	RTM	DW	Sapiens → HM	RTM → HM	DW → HM	Ours*
MJE ↓	29.5	27.5	23.5	20.9	16.9	17.0	**11.8**
$mPCK_{2D}$ ↑	7.7	12.6	14.0	17.4	20.8	20.9	**22.8**
AP ↑	67.3	66.9	66.9	66.9	65.4	66.7	**70.3**

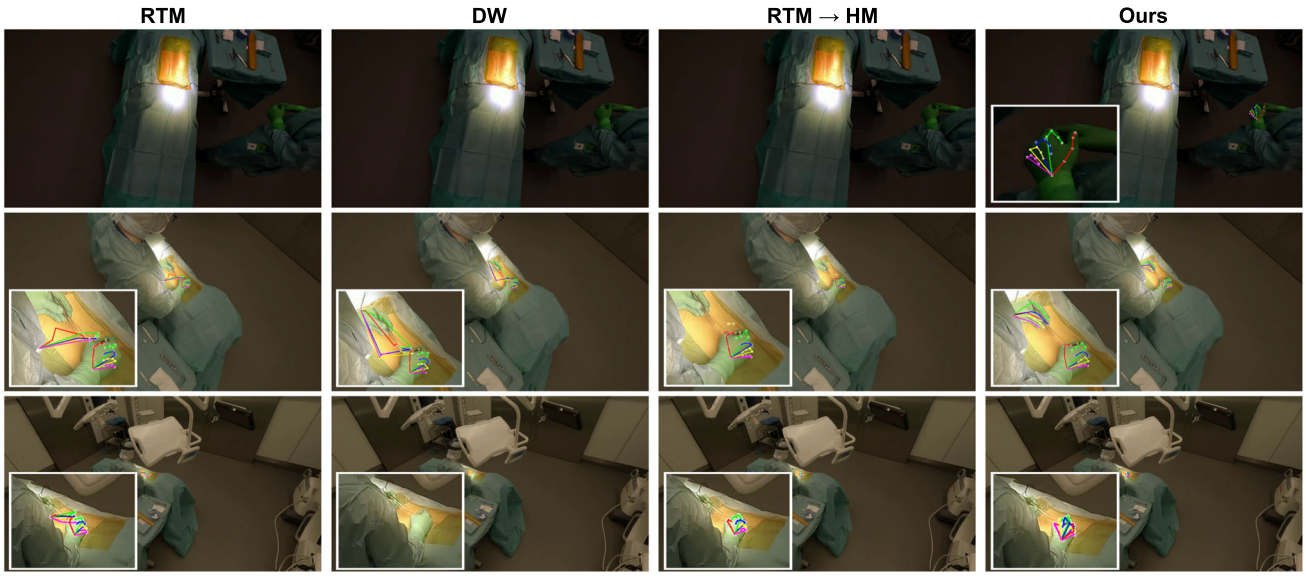


Fig. 3 Qualitative comparison of 2D hand pose predictions. RTM, DW, and RTM → HM baselines are shown alongside our method, which achieves more complete and stable hand poses

Table 2 Quantitative results on the 2D pose predictions in 3D. MPJPE (mm) and MRE (px): lower is better; $mPCK_{2D}$ and $mPCK_{3D}$: higher is better. Results correspond to the 3D optimization in Eq. (1) with weighting factors $\lambda_{\text{reproj}} = 1.0$, $\lambda_{\text{smooth}} = 20.0$, $\lambda_{\text{shape}} = 50.0$, and $\lambda_{\text{biomech}} = 0$ (baseline configuration) for all 2D methods. **Best** and second best

Method	MPJPE ↓	MRE ↓	$mPCK_{2D}$ ↑	$mPCK_{3D}$ ↑
RTM [1]	35.8	39.8	39.8	40.8
RTM → HM [1]	51.9	58.1	42.3	41.4
Ours: RTM + T [22] → HM [1]	**8.5**	**8.2**	**76.6**	**74.7**

Table 3 Ablation study showing the effect of individual optimization terms on reconstruction accuracy. Active components are marked with ✓. The metrics reported are MPJPE and $mPCK_{3D}$ (mm), with L_0–L_2 denoting increasing scene complexity and the mean score over all scenes. **Best** and second best

Keypoint reprojection	✓	✓	✓	✓	✓	✓	✓	✓
Temporal smoothness	×	✓	✓	✓	×	✓	✓	✓
Shape consistency	×	×	✓	✓	×	×	✓	✓
Biomechanical	×	×	×	✓	×	×	×	✓
	MPJPE ↓				$mPCK_{3D}$ ↑			
Complexity level $\mathbf{L}_0$	115.1	**8.0**	**8.0**	24.0	59.5	**75.7**	75.4	71.6
Complexity level $\mathbf{L}_1$	47.6	**8.8**	9.1	14.4	64.5	**74.8**	73.2	64.5
Complexity level $\mathbf{L}_2$	297.9	9.5	**8.8**	69.7	55.9	74.2	**75.4**	23.7
Mean	112.9	**8.4**	8.5	25.9	60.4	**75.1**	74.7	58.6

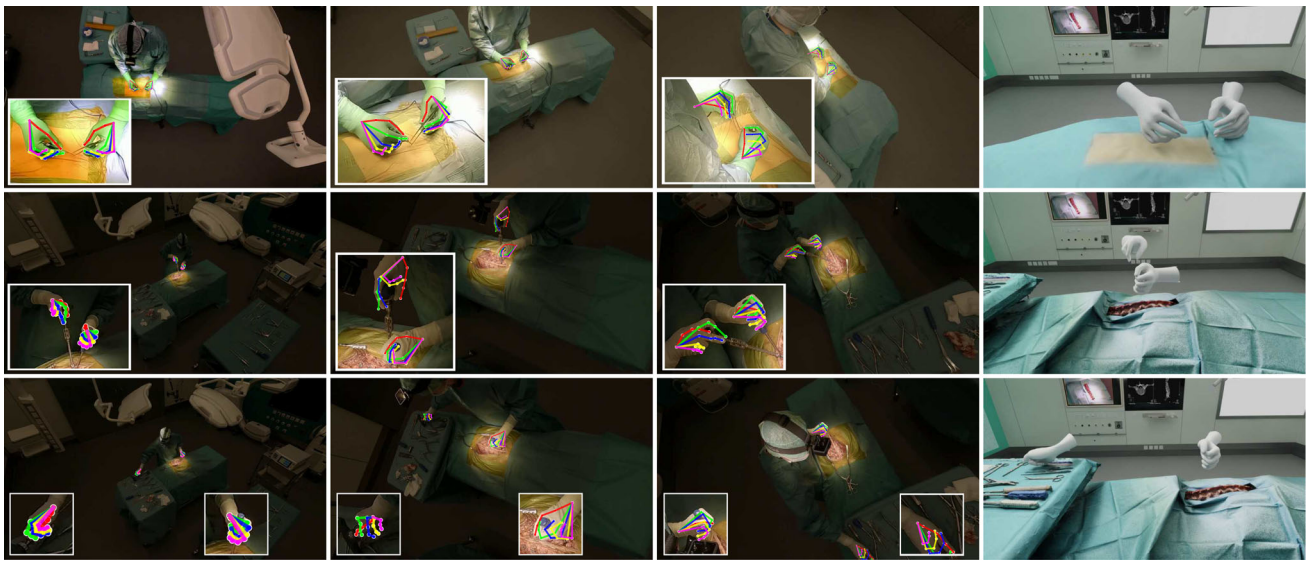


Fig. 4 Qualitative results of the full pipeline showing 3D poses reprojected into the original video streams and a 3D SMPL-X [2] render (right). First row: SPP. Second and third row: SPI

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-026-03610-6.

Acknowledgements The authors would like to thank Jannis Piekarek, main developer of the annotation tool used for the dataset annotations.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. This work has been supported by the OR-X - a swiss national research infrastructure for translational surgery - and associated funding by the University of Zurich and University Hospital Balgrist, and the Innosuisse Flagship project PROFICIENCY No. PFFS-21-19.

Data Availability Code and data required to reproduce the results are available at https://github.com/aelvyr/ORPose.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Approval All procedures were conducted in accordance with the ethical standards of the responsible institutional review body and with the principles of the Declaration of Helsinki and its subsequent amendments (or comparable ethical guidelines). The local ethics committee approved this study (BASEC no. 2021-01196).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Jiang T, Lu P, Zhang L, Ma N, Han R, Lyu C, Li Y, Chen K (2023) Rtmpose: Real-time multi-person pose estimation based on mmpose. arXiv preprint
2. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ (2019) Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10975–10985
3. Zhang M, Cheng X, Copeland D, Desai AD, Guan MY, Brat GA, Yeung S (2020) Using computer vision to automate hand detection and tracking of surgeon movements in videos of open surgery. In: American Medical Informatics Association Annual Symposium (AMIA)
4. Bkheet E, D'Angelo A-L, Goldbraikh A, Laufer S (2023) Using hand pose estimation to automate open surgery training feedback. Int J Comput Assist Radiol Surg 18(7):1279–1285
5. Hein J, Seibold M, Bogo F, Farshad M, Pollefeys M, Fürnstahl P, Navab N (2021) Towards markerless surgical tool and hand pose estimation. Int J Comput Assist Radiol Surg 16:799–808
6. Müller L-R, Petersen J, Yamlahi A, Wise P, Adler TJ, Seitel A, Kowalewski K-F, Müller B, Kenngott H, Nickel F, Maier-Hein L (2022) Robust hand tracking for surgical telestration. Int J Comput Assist Radiol Surg 17(8):1477–1486. https://doi.org/10.1007/s11548-022-02637-9
7. Zhang X, Wang J, Dai X, Shen S, Chen X (2023) A non-contact interactive system for multimodal surgical robots based on leap-motion and visual tags. Front Neurosci 17:1287053
8. Wang J, Zhang X, Chen X, Song Z (2023) A touch-free human-robot collaborative surgical navigation system based on hand gesture recognition. Front Neurosci 17:1200576
9. Zimmermann C, Brox T (2017) Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV)

10. Iqbal U, Molchanov P, Breuel T, Gall J, Kautz J (2018) Hand pose estimation via latent 2.5d heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 118–134
11. Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C (2018) Ganerated hands for real-time 3d hand tracking from monocular rgb. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 49–59
12. Xie P, Xu W, Tang T, Yu Z, Lu C (2024) Ms-mano: Enabling hand pose tracking with biomechanical constraints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2382–2392
13. Hewitt C, Saleh F, Aliakbarian S, Petikam L, Rezaeifar S, Florentin L, Hosenie Z, Cashman TJ, Valentin J, Cosker D, Baltrusaitis T (2024) Look ma, no markers: Holistic performance capture without the hassle. ACM Trans Graph 43(6):1–12. https://doi.org/10.1145/3687772
14. Avola D, Cinque L, Fagioli A, Foresti GL, Fragomeni A, Pannone D (2022) 3d hand pose and shape estimation from rgb images for keypoint-based hand gesture recognition. Pattern Recogn 129:108762. https://doi.org/10.1016/j.patcog.2022.108762
15. Spurr A, Iqbal U, Molchanov P, Hilliges O, Kautz J (2020) Weakly supervised 3d hand pose estimation via biomechanical constraints. In: European Conference on Computer Vision (ECCV)
16. Khirodkar R, Bagautdinov T, Martinez J, Su Z, James A, Selednik P, Anderson S, Saito S (2024) Sapiens: Foundation for human vision models. European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-031-73235-5_12
17. Zimmermann C, Ceylan D, Yang J, Russell B, Argus M, Brox T (2019) Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: IEEE International Conference on Computer Vision (ICCV)
18. Zimmermann C, Argus M, Brox T (2021) Contrastive representation learning for hand shape estimation. In: Pattern Recognition. DAGM GCPR
19. Yang Z, Zeng A, Yuan C, Li Y (2023) Effective whole-body pose estimation with two-stages distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). CV4Metaverse Workshop
20. Srivastav V, Gangi A, Padoy N (2022) Unsupervised domain adaptation for clinician pose estimation and instance segmentation in the operating room. Med Image Anal 80:102525. https://doi.org/10.1016/j.media.2022.102525
21. Jocher G, Qiu J Ultralytics YOLO11
22. Xiong Y, Zhou C, Xiang X, Wu L, Zhu C, Liu Z, Suri S, Varadarajan B, Akula R, Iandola F, Krishnamoorthi R, Soran B, Chandra V (2025) Efficient track anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
23. Meyer J, Giraud F, Wüthrich J, Fürnstahl P, Calvet L (2025) Rocsync: Millisecond-accurate temporal synchronization for heterogeneous camera systems. arXiv preprint
24. Flückiger T, Hein J, Fischer V, Fürnstahl P, Calvet L (2025) Automatic calibration of a multi-camera system with limited overlapping fields of view for 3d surgical scene reconstruction. International Journal of Computer Assisted Radiology and Surgery
25. Lin T-Y, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollár P (2014) Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-319-10602-1_48

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.