International Journal of Computer Assisted Radiology and Surgery (2023) 18:117–125
https://doi.org/10.1007/s11548-022-02761-6

ORIGINAL ARTICLE

# Temporally guided articulated hand pose tracking in surgical videos

Nathan Louis[1] · Luowei Zhou[2] · Steven J. Yule[3] · Roger D. Dias[4] · Milisa Manojlovich[5] · Francis D. Pagani[6] · Donald S. Likosky[6] · Jason J. Corso[1]

Received: 17 March 2022 / Accepted: 13 September 2022 / Published online: 3 October 2022


## Abstract

**Purpose** Articulated hand pose tracking is an under-explored problem that carries the potential for use in an extensive number of applications, especially in the medical domain. With a robust and accurate tracking system on surgical videos, the motion dynamics and movement patterns of the hands can be captured and analyzed for many rich tasks.

**Methods** In this work, we propose a novel hand pose estimation model, **CondPose**, which improves detection and tracking accuracy by incorporating a pose prior into its prediction. We show improvements over state-of-the-art methods which provide frame-wise independent predictions, by following a temporally guided approach that effectively leverages past predictions.

**Results** We collect *Surgical Hands*, the first dataset that provides multi-instance articulated hand pose annotations for videos. Our dataset provides over 8.1k annotated hand poses from publicly available surgical videos and bounding boxes, pose annotations, and tracking IDs to enable multi-instance tracking. When evaluated on *Surgical Hands*, we show our method outperforms the state-of-the-art approach using mean Average Precision, to measure pose estimation accuracy, and Multiple Object Tracking Accuracy, to assess pose tracking performance.

**Conclusion** In comparison to a frame-wise independent strategy, we show greater performance in detecting and tracking hand poses and more substantial impact on localization accuracy. This has positive implications in generating more accurate representations of hands in the scene to be used for targeted downstream tasks.

**Keywords** Articulated pose · Surgical videos · Computer vision · Hand pose · Video tracking

## Introduction

Machine learning and computer vision have become increasingly integrated with healthcare in the medical community. This is apparent in the myriad of tasks, such as tumor segmentation [1], technical skill assessment [2–6], and tool detection and tracking [7–10]. Here we study the problem of articulated hand pose tracking in the surgical domain. Tracking hand poses can facilitate other useful tasks, such as technical skill assessment, temporal action recognition, and training surgical residents. Pose tracking in the computer vision com-

✉ Nathan Louis
natlouis@umich.edu

✉ Jason J. Corso
jjcorso@umich.edu

Luowei Zhou
luozhou@microsoft.com

Steven J. Yule
steven.yule@ed.ac.uk

Roger D. Dias
rdias@bwh.harvard.edu

Milisa Manojlovich
mmanojlo@med.umich.edu

Francis D. Pagani
fpagani@med.umich.edu

Donald S. Likosky
likosky@med.umich.edu

1 EECS, University of Michigan, Ann Arbor, MI, USA

2 Cloud and AI, Microsoft, Redmond, WA, USA

3 Clinical Surgery, University of Edinburgh, Edinburgh, Scotland, UK

4 Emergency Medicine, Harvard Medical School, Boston, MA, USA

5 School of Nursing, University of Michigan, Ann Arbor, MI, USA

6 Cardiac Surgery, University of Michigan, Ann Arbor, MI, USA

munity is primarily centered around human poses [11–19], while medical works focus on detecting and tracking surgical instruments [7–10]. Tracking surgical instruments is useful but these instruments are inherent to the surgical procedures seen during training. Instead we abstract away the emphasis on surgical instruments where articulated hand tracking will be more applicable to broad surgical tasks. Articulated hand pose tracking can highlight important properties such as grip, motion, and tension that human experts often attend to when evaluating videos.

A challenge in pose tracking is the temporal consistency of predictions between frames, the lack of which leads to flickering and improbable changes in estimated poses. Existing works [11,14,17–19] in articulated pose tracking use frame-wise independent predictions along with post-processing when tracking [12,13,15,16] to gather temporal context. However, they do not integrate past inferences when localizing joints. We address this by proposing **CondPose**, a new model that performs predictions conditioned on the pose estimates from prior frames. In Fig. 1, we show a comparison of both approaches: the baseline using frame-wise independent predictions and our model using conditional predictions. The initial estimate may fluctuate due to varying factors, such as lighting, hand orientation, or motion blur. But we find that using prior predictions as guidance, we can improve our localization accuracy. The internal representation of this object's state (position, appearance, and classification) is a function of its current state and previous states. By learning this Markovian prior for the prediction of hand joints, we can improve both pose estimation and consequently tracking accuracy.

There is a lack of data and benchmarks for articulated hand pose tracking. To address this, we collect a novel dataset featuring intra-operative videos of real surgeries, *Surgical Hands*. We annotate the articulated hand poses of surgeons which subsumes both surgical instrument and non-instrument actions, e.g., suturing, knot-tying, and gesturing. We are, to the best of our knowledge, the first to introduce a labeled dataset for both detection and tracking of multiple articulated hand poses. We benchmark our dataset against existing tracking baselines and demonstrate the superiority of our proposed approach on both hand pose estimation and tracking.

Our contributions are as follows:

- We introduce **CondPose**, a novel deep network that takes advantage of confident prior predictions to improve localization accuracy and tracking consistency.
- We present *Surgical Hands*[1], a new video dataset for multi-instance articulated hand pose estimation and tracking in the surgical domain.
- We set new state-of-the-art benchmark performance on *Surgical Hands*.

[1] Both the code and dataset are available at https://github.com/MichiganCOG/Surgical_Hands_RELEASE.

## Related works

### Articulated pose estimation and tracking

#### Surgical instruments

Data-driven methods in the medical video domain primarily involve RAS videos. Works in this space [3–5] traditionally use kinematic data directly, requiring an external apparatus to capture these measurements. But full kinematic information is only available for robotic-controlled tools, even less so for hand-held instruments. Adding any external apparatus to capture kinematic data can negatively impact the costs, flexibility, and performance of certain operations. For detection, pure computer vision-based approaches extract information directly from video data to perform object detection. Many vision works use a region proposal network to perform localization [7,20,21], segmentation [9,22], and articulated pose estimation [8,23] from images.

To incorporate tracking, existing works may use a similarity function based on weighted mutual information [24] or Bayesian filtering as part of a minimization problem [25]. Nwoye et al. [10] are the first to measure the Multiple Object Tracking Accuracy (MOTA) [26] for surgical instruments in this setting, using a weakly-supervised approach with coarse binary labels indicating the presence or absence of seven surgical instruments. However, their evaluation contains at most one unique type of tool at each frame; hence, can be narrowed down to an object detection problem. Unlike their work, we track multiple instances of the same object in each frame. We also use MOTA as part of our benchmark when tracking hands in our videos.

#### Human pose

Pose estimation and tracking is commonly applied to images and videos of people, grouped into top-down [12–16] and bottom-up [17–19] strategies. Top-down methods detect all persons from an image, then regress each human pose independently using a pose estimation network. Bottom-up methods detect all joints in an image, and use bipartite matching and graph minimization techniques to assign joints to each person. As top-down approaches typically perform best in practice, we follow this paradigm. For tracking, [12] uses a greedy matching from IoU (intersection-over-union) overlap and optical flow to propagate bounding boxes between frames, [13] use deformable convolutions to warp predictions between frames, and [15] introduce a Graph Convolutional Network (GCN) [27] to match learned embeddings between

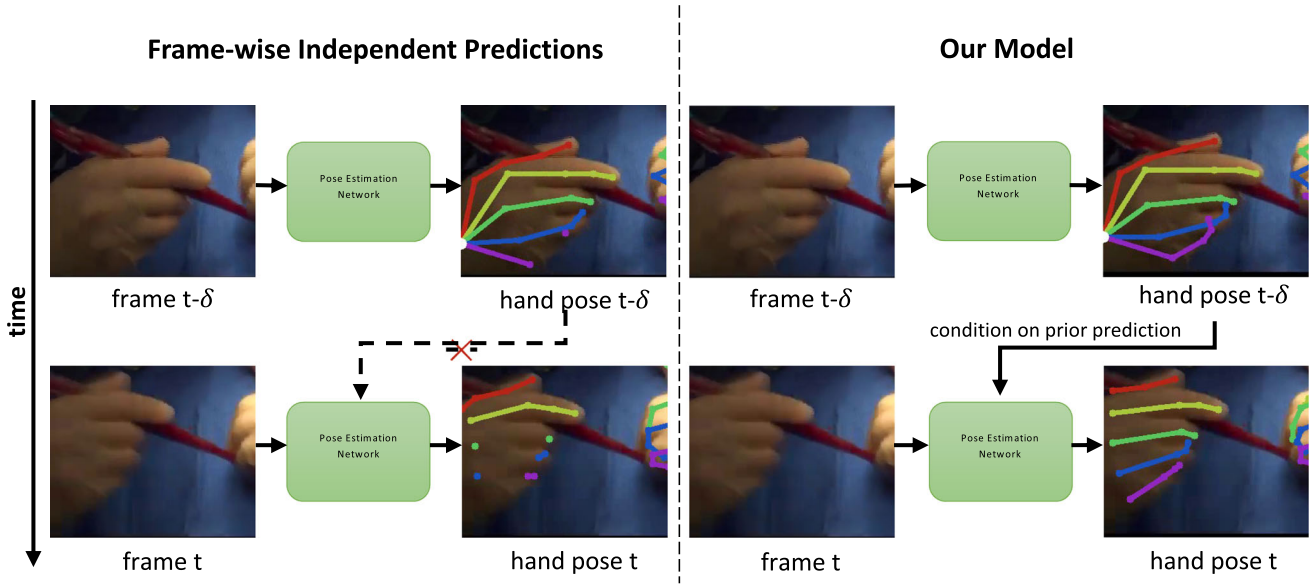


**Fig. 1** On the left, a method only performing frame-wise independent predictions may miss out on properly localizing joints, while on the right, temporally passing past predictions from previous frames improves the network's localization

human poses. A GCN is a neural network whose input consists of a set of nodes and edges, performing convolution operations on the relations of nodes. The inherent structure of this graph can improve quality of learned features as well as abstracting from limitations of a 2D space. These approaches spatially shift pose predictions, which cannot overcome certain factors (e.g., missed detections). In contrast, we address this problem at the detection step by integrating past pose observation(s) into each new predicted output.

### Hand pose

Current works on 2D hand pose estimation [28–30] are analogous to human pose estimation. Zhang et al. [31] performs pose tracking, using a disparity map from stereo camera inputs to estimate a 3D hand pose. However their data consists of only a single subject's hand and at most one detection per frame. There are many image datasets [28,30–32] for hand pose estimation, from a combination of manual, synthetic, and predicted annotations. But none satisfy the conditions of multiple object instances and tracking from video, more so in a surgical setting. Therefore, we introduce the *Surgical Hands* dataset for multi-instance articulated hand pose tracking. Our dataset includes varying lighting conditions, fast movement, and diversity in scene appearances. Distinctively, we also include gloved hands, which appear in contrasting colors such as latex and green.

## Method

We propose **CondPose**, to perform articulated pose detection and tracking by incorporating previous observations as prior guidance. We show our model in Fig. 2. While the baseline produces a heatmap from each hand using a pose estimation network, we leverage past predictions to produce conditioned hand pose outputs, improving detection performance during inference. While we design **CondPose** with video data in mind, we begin with pretraining on image data, finetuning on our video dataset, *Surgical Hands*, and lastly, comparing between different tracking methods.

### Hand pose estimation in images

We first pretrain on image data, defining the input and output for the pose estimation network, $P$, as $\hat{\mathcal{H}} = P(\mathcal{I})$. The input is an image crop $\mathcal{I}$, $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, and the output is a predicted heatmap $\hat{\mathcal{H}}$, $\hat{\mathcal{H}} \in \mathbb{R}^{H' \times W' \times J}$. Here $H$, $W$ represents the input image height and width and $H'$, $W'$ are the output heatmap height and widths. $J$ represents the number of predicted joints of each hand. Each image crop is scaled to 2.2 times the total area of the hand bounding box. We train using the mean squared error (MSE) between the ground truth and predicted heatmaps as $\mathcal{L} = \|(\mathcal{H} - \hat{\mathcal{H}}) \odot \mathcal{M}\|^2$. The ground truth heatmaps, $\mathcal{H}$, are generated from 2D Gaussians centered on each annotated keypoint. $\mathcal{M}$, is included to

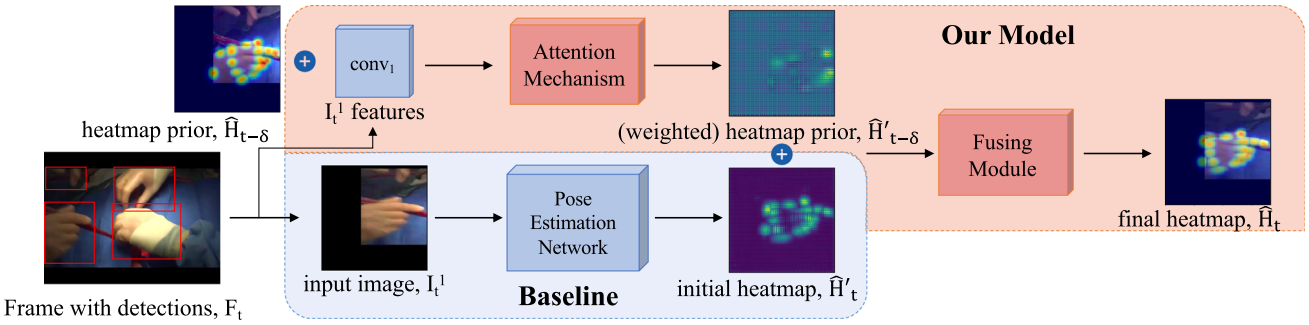


**Fig. 2** The baseline generates a heatmap, $\hat{\mathcal{H}}'_t$, for each detection using a pose estimation network. In our model, we provide additional information by incorporating a heatmap prior from $t-\delta$. Concatenating the image features at $t$ with $\hat{\mathcal{H}}_{t-\delta}$, we pass this through our attention mechanism to produce a weighted heatmap prior, $\hat{\mathcal{H}}'_{t-\delta}$. Both $\hat{\mathcal{H}}'_t$ and $\hat{\mathcal{H}}'_{t-\delta}$ are concatenated and passed through the fusing module, using context from both heatmaps to produce the final articulated hand pose. (The initial and final heatmaps represent real outputs from the network, while the heatmap prior (during training) shows ground truth at $t-\delta$)

mask out un-annotated joints. The output joint locations are the max value positions in the third channel of $\hat{\mathcal{H}}$. After pretraining, we finetune our model on videos to learn conditional hand pose predictions.

## Hand pose estimation in videos

While image data cannot be used to learn our conditional hand pose predictions, we can initialize weights to speed up our training process and improve generalizability. We finetune **CondPose** on *Surgical Hands*, shown in the top portion of Fig. 2. To incorporate a prior branch, we introduce a heatmap prior, $\hat{\mathcal{H}}_{t-\delta}$, a pose estimate of the same object from $t-\delta$. Our model performs conditional predictions, defined as

$$\hat{\mathcal{H}}_t = M_{\text{fus}}(P(\mathcal{I}_t); M_{\text{att}}(v_t; \hat{\mathcal{H}}_{t-\delta})). \quad (1)$$

In contrast to our previous definition of $P$, $\hat{\mathcal{H}}_t$ is now conditioned on predictions at a previous time step $t-\delta$. Our model is further composed of two branches: the attention mechanism, $M_{\text{att}}$, and the fusing module, $M_{\text{fus}}$. $M_{\text{att}}$ contextualizes the prior heatmap prediction, $\hat{\mathcal{H}}_{t-\delta}$, with image features, $v_t$ ($conv\_1$ in our experiments), at time $t$. This branch relates the visual representation and the localized heatmap prior, ideally learning to weight each joint prior accordingly. $M_{fus}$ produces a merged final heatmap from the initial prediction, $\hat{\mathcal{H}}'_t$, and weighted heatmap prior, $\hat{\mathcal{H}}'_{t-\delta}$. $M_{att}$ and $M_{fus}$ are both composed of two convolutional layers, followed by transposed convolution, with ReLU nonlinearities in-between.

During training the prior is selected from frame $t-\delta$. If the object does not exist at that frame, we use earlier frames up until the first occurrence. If a corresponding object does not exist on any previous frames, then the prior, $\hat{\mathcal{H}}_{t-\delta}$, is set as a zeros heatmap. This is expected behavior during evaluation, because priors do not yet exist at frame one. Also during evaluation, unlike training, the prior associated with the current detection is unknown. Given $n$ priors from time $t-1$, $\{\hat{\mathcal{H}}^1_{t-1}, \hat{\mathcal{H}}^2_{t-1}, \ldots \hat{\mathcal{H}}^n_{t-1}\}$, and $k$ detections at time $t$, $\{\hat{\mathcal{I}}^1_{t-1}, \hat{\mathcal{I}}^2_{t-1}, \ldots \hat{\mathcal{I}}^k_{t-1}\}$ we pass all pairs through the network to generate candidates. The heatmap with the highest average confidence score is selected as the output for that detection.

## Matching strategies for tracking

Following the detect-then-track paradigm, we require a matching strategy to performing tracking. Given $n$ hands at time $t-1$ and $m$ hands at time $t$ we use a similarity function to derive similarity measures between each pair at $t-1$ and $t$. Common methods are intersection-over-union (IoU) of bounding boxes, average L2-distance of the predicted joint locations, or L2-distance between the graph pose embeddings. Similar to Ning et al. [15] we train a GCN to output the embedding of each input hand pose, $\mathcal{X}$, defined simply as $\hat{p} = GCN(\mathcal{X})$. Here $\mathcal{X} \in \mathbb{R}^{J \times C}$, where $J$ is the number of joints and $C$ is the number of channels. For training, we use the contrastive loss [33], $\mathcal{L} = \frac{1}{2}\left(y * d + (1-y) * max\left(0, (m-d)^2\right)\right)$. The contrastive loss places embeddings close in perceptual distance. For a pair of embeddings $\hat{p}^1_v$ and $\hat{p}^2_v$, the variable $d$ represents the L2-distance between the two, $d = \|\hat{p}^1_v - \hat{p}^2_v\|^2$. $y$ is a binary label indicating the same hand, 1, or different hands, 0. $m$ is the margin variable, a hyperparameter used for tuning. For each item in our minibatch, positive pairs are selected between adjacent frames with probability $p = 0.5$ and negative pairs are selected from the same video with $p = 0.4$ or from a different video with $p = 0.1$. We evaluate our trained GCN models using the classification accuracy between pairs of selected hands, achieving classification accuracies of $> 97\%$.

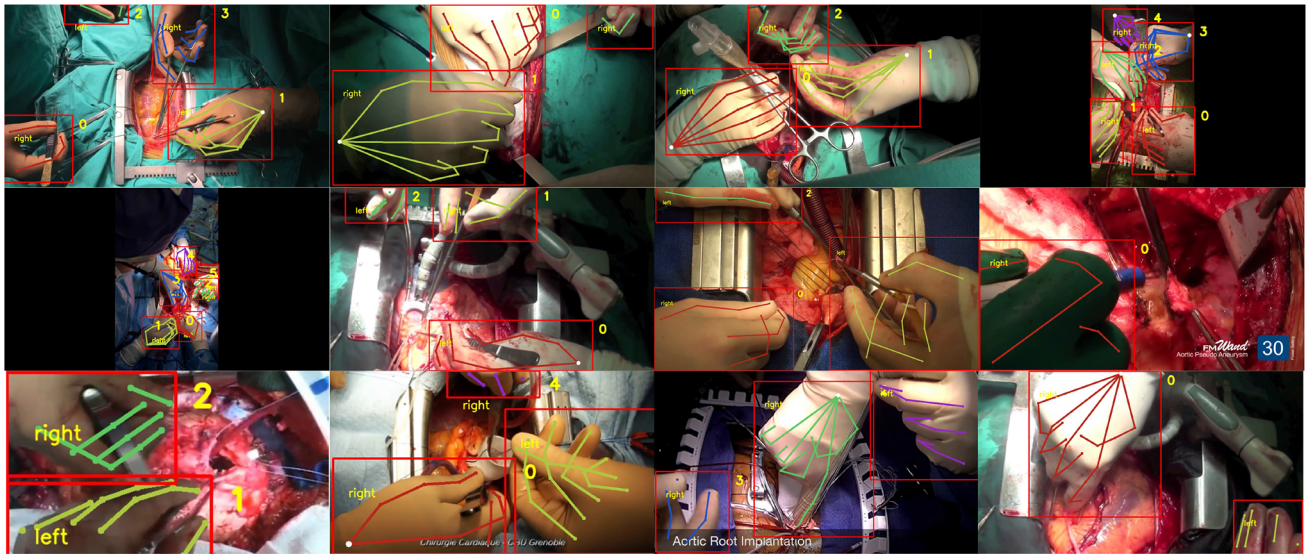


**Fig. 3** We show samples from our annotations. Each hand is labeled with a bounding box, handedness, tracking id, and visibility of joints

## Dataset

We lack data for training and benchmarking models on multi-instance hand tracking. Therefore we introduce *Surgical Hands*, a novel video dataset for multi-instance articulated hand pose estimation and tracking in the surgical domain, the first of its kind. From publicly available data, we collect 28 videos with a view of the hands of surgical team members during the operation. From those videos, we extract 76 clips sampled at 8 frames per second and collect bounding box, class label, tracking id, and pose annotations using Amazon Mechanical Turk (AMT) and a modified version of Visipedia Annotation Tools. [2] We show samples of our annotations in Fig. 3. Each hand is labeled with the handedness (left/right), 21 joints, and properties for each joint: visible, occluded or non-available. Visible implies that the joint is visibly on screen, occluded means the joint is obstructed but its position can be estimated, not-available means the joint position cannot be inferred or it is off-screen. From our collected data, we have a total 2, 838 annotated frames and 8, 178 unique hand annotations from 21 unique annotators. Each annotated frame contains a mean of 2.88 hands, median of 3 hands, and a maximum of 7 hands.

## Experiments and evaluation

### Implementation details

We adopt a ResNet-152 pose estimation model [12] to first train on hand pose image data, CMU Manual Hands and Synthetic Hands [28]. We use a batch size of 16, training for 30 epochs, with an Adam optimizer and a learning rate of $1e^{-3}$. When finetuning on *Surgical Hands* we use leave-one-out cross-validation and split our data into 28 different folds. Clips belonging to the same video are in the same validation fold, and the reported metrics are averaged across all folds. We employ a variant of curriculum learning that gradually transitions to predicted priors from ground truth priors. A predicted prior at $t - \delta$ is sampled with a probability of $p = 0.10 * epoch$, until only predictions are used for training at epoch 10 and onward. We empirically select $\delta = 3$ during training. For all training, we apply random rotations and horizontal flipping as data augmentation. When training the GCN for tracking, we using a batch size of 32 and train for 60 epochs and an initial learning rate of $1e^{-3}$. We normalize $\mathcal{X}$ to 0-1, relative to keypoint positions along the bounding box. The input dimension for each input is $J \times C$ where $J$ represents the number of joints and $C$ is the number of channels. We use $C = 2$ for x-y coordinates and $C = 3$ to include annotation state (0 = unannotated, 1 = annotated, or 0-1 for predicted keypoints). We adopt a two-layer Spatio-Temporal GCN [15,34] to output a 128-dimensional embedding of each pose.

### Detection performance

We evaluate detection performance using mean Average Precision (mAP), the choice metric in human pose evaluation, on our *Surgical Hands* dataset. MAP is computed using the Probability of Correct Keypoints (PCK), measuring the probability of correctly localizing keypoints within a normalized threshold distance, $\sigma$. This threshold distance, $\sigma$=0.2, is empirically chosen to be roughly the ratio between the length

[2] https://github.com/visipedia/annotation_tools.

**Table 1** Mean Average Precision (mAP)

| Model | Wrist | Thumb | Index | Middle | Ring | Pinky | mAP |
|---|---|---|---|---|---|---|---|
| Baseline [12] | **67.23** | 60.12 | 63.29 | 53.77 | 48.29 | 39.28 | 53.59 |
| Our model | 65.51 | **62.66** | **64.99** | **57.88** | **51.40** | **44.26** | **56.66** |

Performance is averaged across all folds
The bolded numbers represent the best performing scores, in comparison between method

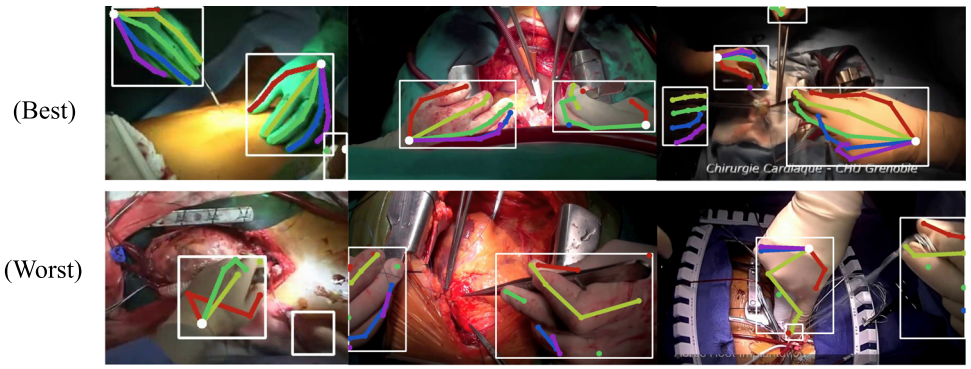


**Fig. 4** We show qualitative samples of frames from the best performing (top row) and lower performing (bottom row) videos. (Best viewed in color)

**Table 2** We optimize for the multiple Object Tracking Accuracy (MOTA), each performance metric is averaged across all validation folds

| Model | MOTA wrist | MOTA thumb | MOTA index | MOTA middle | MOTA ring | MOTA pinky | MOTA total | MOTP total | Prec. total | Rec. total | $F_1$ Score total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline [12] | **36.7** | **45.83** | 57.35 | 45.53 | 34.63 | 8.49 | 38.27 | 85 | **78.3** | 59.13 | 67.37 |
| Our model | 30.99 | 44.74 | **58.21** | **48.90** | **36.46** | **10.39** | **39.31** | **85.28** | 77.61 | **62.69** | **69.35** |

The bolded numbers represent the best performing scores, in comparison between method

of a thumb joint and the enclosing bounding box. Pose predictions are matched to ground truth poses based on the highest PCK and unassigned predictions are counted as false positives. AP for each joint is computed and mAP is reported across the entire dataset. In Table 1 we report the mAP at the highest MOTA score (defined in the next section) for each model. With our recursive heatmap strategy we are able to obtain higher average precision across the different joints in the hand. In Fig. 4 we show qualitative examples of our hand pose estimation on various frames from our *Surgical Hands* dataset. The top row clips are sampled from the best performing clips, while the bottom row are from the worst performing clips. We see that the model suffers most in cases of heavy occlusion, where the camera view excludes the majority of the hand. Ambiguity in the position of the hand furthers the localization errors, e.g., top-down view with most fingers occluded. The best performing cases are those with balanced lighting and an unambiguous view of the first few digits.

## Tracking performance

To measure tracking performance, we use Multiple Object Tracking Accuracy (MOTA) which also takes into account the consistency of localized keypoints between frames. MOTA [26] is defined as:

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t G_t} \tag{2}$$

This encapsulates errors that may occur during multiple object tracking: false negatives (FN), false positives (FP), and identity switches (IDSW). FN are joints for which no hypothesis/prediction was given, FP are the hypothesis for which no real joints exists, and IDSW are occurrences where the tracking id for two joints are swapped. $G$ represents the total number of ground truth joints. The range of values for the MOTA score is $(-\infty$ to 100].

We measure perform tracking using three methods: IoU, $L2$-distance, and GCN. Intersection-over-union (IoU) measures overlap of two bounding boxes using the ratio: area of intersection over total area, between subsequent frames in our case. $L2$-distance measures the average $L2$ distance of regressed keypoints between frames. GCN measures the embedding similarity between the encoded keypoints to determine matches. We show quantitative results from our experiments in Table 3 and the per-joint performance in Table 2. Each row is maximized for the highest MOTA score across all hyperparameters, shown along with its corresponding mAP. Our method has a higher MOTA score across all of the videos, but our corresponding mAP scores are greater

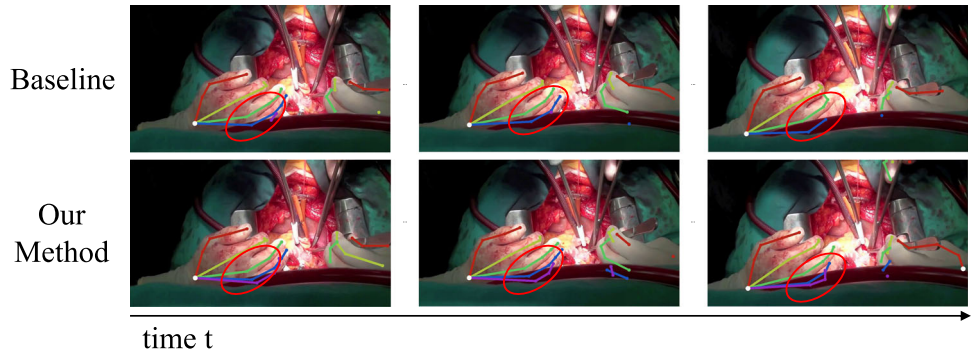


**Fig. 5** We show a qualitative comparison between the baseline model and our method. We note a higher recall and consistency between frames, as shown for the hand to the left. Even when the pinky finger is not visible, the past predictions reinforces those joint locations

**Table 3** MOTA performance between matching strategies, averaged across all folds. Each row is optimized for highest MOTA performance. Matching strategies share the same base model, so it is possible for them to share the same mAP score

| Model | Matching strategy | Perfect det. | | Object det. | |
|---|---|---|---|---|---|
| | | mAP | MOTA | mAP | MOTA |
| Baseline [12] | IoU | 53.59 | 38.27 | 48.15 | 31.46 |
| | *L2* | 52.65 | 37.78 | 47.44 | 31.14 |
| | GCN | 52.65 | 36.78 | 47.44 | 30.03 |
| Our model | IoU | **56.66** | **39.31** | **50.04** | **33.19** |
| | *L2* | 56.66 | 38.94 | 50.04 | 32.84 |
| | GCN | 56.66 | 38.22 | 50.04 | 32.24 |

The bolded numbers represent the best performing scores, in comparison between method

**Table 4** Ablation analysis using IoU matching strategy ($\delta = 1$)

| Model variant | Matching strategy | Perfect det. | |
|---|---|---|---|
| | | mAP | MOTA |
| NC-NA | IoU | 55.23 | 38.31 |
| NC | IoU | 56.00 | 38.13 |
| NA | IoU | 54.70 | 38.45 |
| Full model | IoU | **56.66** | **39.31** |

*NC* No convolutional feature map, *NA* No attention mechanism
The bolded numbers represent the best performing scores, in comparison between method

**Table 5** Effect of $\delta$

| Model variant | Matching strategy | Perfect det. | |
|---|---|---|---|
| | | mAP | MOTA |
| $\delta = 1$ | IoU | **58.64** | 39.03 |
| $\delta = 2$ | IoU | 54.71 | 38.42 |
| $\delta = 3$ | IoU | 56.66 | **39.31** |
| $\delta = 4$ | IoU | 56.35 | 38.09 |

Each model is trained with a separate $\delta$ value
The bolded numbers represent the best performing scores, in comparison between method

by a much larger margin. This points to our advantage from temporally leveraging predictions from previous frames during the detection step. We show an example in Fig. 5, in a frame-by-frame comparison between the baseline and our method, we note a higher recall and improved localization. While the last digit is obstructed, its position can be reasonably inferred. In the last two columns of Table 3 we use an object detector to detect hands, the prior two columns (perfect detections) use the manual annotations. Training an object detector on 100 Days of Hands (100DOH) [35], we see a lower localization and tracking accuracy but a consistent trend from the baseline. The quality of the detections serve as a bottleneck, but the margins of improvements are very similar. While trained with perfect detections as priors, they are not required to maintain performance in practice.

## Ablation analysis

We perform an ablative analysis on the convolutional map in $M_{\text{att}}$ and the fusing module $M_{\text{fus}}$. We experiment with no prior convolutional feature map (NC), no attention mechanism (NA), and removal of both (NC-NA), showing our results in Table 4. Our full model has the highest scores overall. The attention mechanism and convolutional feature maps have opposing effects on the mAP and MOTA scores. The NC model does not use a convolutional feature map from frame $t$, so the fusing module is applied directly to both un-altered heatmaps from $t - \delta$ and $t$. We found this increases the mAP value, but lowers the MOTA score. The NA model directly concatenates the convolutional features and the heatmaps, with no attention mechanism. This has the opposite effect, decreasing the mAP significantly but slightly increasing the overall MOTA score. Without contextual convolutional features (NC and NC-NA), the model can still learn to use the

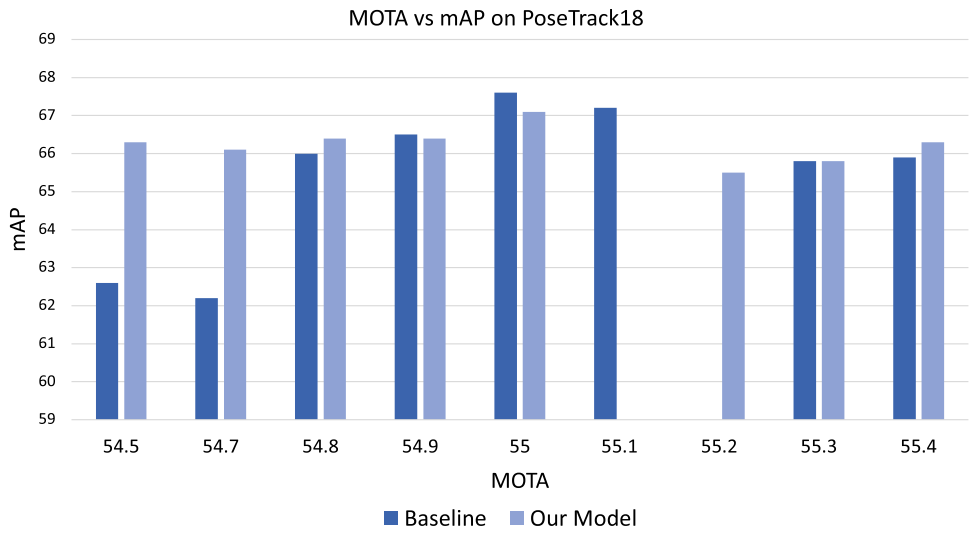


**Fig. 6** Optimized for maximum MOTA score, we show the top performing models on PoseTrack18. Consistent with our earlier findings, our model maintains a higher mAP for comparable MOTA scores

prior prediction and improve its detection score. On the contrary, no attention mechanism brings a drop in mAP, which may be attributed to an unrefined prior with noisy features. The small increase in the MOTA score is likely from fewer false positives produced by that model, due to a slightly lower mAP.

We also explore the value of our hyperparameter, $\delta$, during training. We use values $\delta = \{1, 2, 3, 4\}$ and show our results in Table 5. Optimizing for highest MOTA score, we found $\delta = 3$ to be best with 39.31, followed by $\delta = 1$ with a smaller MOTA score (39.03) but a higher mAP (58.64 vs 56.66). We find a nonlinear correlation between the mAP and MOTA scores, showing a trade-off in mAP when optimizing for the tracking performance. The best strategy is one that maximizes MOTA accuracy with minimal loss in localization precision.

### Evaluation on human pose

We executed additional experiments on the PoseTrack18 dataset between our model and our re-implementation of the baseline. From Fig. 6, we show a narrowed gap in performance but our findings are consistent with our earlier experiments. Our model maintains a higher mAP score for the highest MOTA values. Given the trade-off that occurs between mAP and MOTA, this means our model is more likely to retain its localization precision at higher tracking accuracies.

## Conclusion

In this work, we introduce *Surgical Hands*, the first articulated multi-hand pose tracking dataset of its kind. Additionally we introduce **CondPose**, a novel network that makes conditional hand pose predictions by incorporating past observations as priors. We show that when compared with a frame-wise independent strategy, we have better performance in localizing and tracking hand poses. More so, a higher localization accuracy for comparable tracking performance. While tracking drives the consistency of joints through time, the actual shape and characteristics of the hand is described by the localization precision. With a higher localization precision and better tracking still, we can guarantee a better representation of the hands in the scene. While not the focus of this work a reliable hand tracking method can provide a salient signal that can be used to approximate surgical skill or understanding actions.

**Funding** This project was supported by the National Heart, Lung, and Blood Institute (NHLBI: R01HL146619) and the University of Michigan (U-M's Mcubed Program). Opinions expressed in this manuscript do not represent those of The NIH or the US Department of Health and Human Services or the US Department of Veterans Affairs.

## Declarations

**Conflicts of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

**Informed consent** This article does not contain patient data.

## References


1. Malathi M, Sinthia P (2019) Brain tumour segmentation using convolutional neural network with tensor flow. Asian Pac J Cancer Prev: APJCP 20(7):2095
2. Dias RD, Gupta A, Yule SJ (2019) Using machine learning to assess physician competence: a systematic review. Acad Med 94(3):427–439
3. Tao L, Elhamifar E, Khudanpur S, Hager GD, Vidal R (2012) Sparse hidden markov models for surgical gesture classification and skill evaluation. In: international conference on information processing in computer-assisted interventions. Springer, pp 167–177
4. Zappella L, Béjar B, Hager G, Vidal R (2013) Surgical gesture classification from video and kinematic data. Med Image Anal 17(7):732–745
5. Forestier G, Petitjean F, Senin P, Despinoy F, Huaulmé A, Fawaz HI, Weber J, Idoumghar L, Muller P-A, Jannin P (2018) Surgical motion analysis using discriminative interpretable patterns. Artif Intell Med 91:3–11
6. Kumar S, Ahmidi N, Hager G, Singhal P, Corso J, Krovi V (2015) Surgical performance assessment. Mech Eng 137(09):7–10
7. Sarikaya D, Corso JJ, Guru KA (2017) Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. IEEE TMI 36(7):1542–1549
8. Colleoni E, Moccia S, Du X, De Momi E, Stoyanov D (2019) Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers. IEEE Robot Autom Lett 4(3):2714–2721
9. Ni Z-L, Bian G-B, Xie X-L, Hou Z-G, Zhou X-H, Zhou Y-J (2019) Rasnet: segmentation for tracking surgical instruments in surgical videos using refined attention segmentation network. In: 2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 5735–5738
10. Nwoye CI, Mutter D, Marescaux J, Padoy N (2019) Weakly supervised convolutional lstm approach for tool tracking in laparoscopic videos. IJCARS 14(6):1059–1067
11. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B (2018) Posetrack: a benchmark for human pose estimation and tracking. In: IEEE CVPR, pp 5167–5176
12. Xiao B, Wu H, Wei Y (2018) Simple baselines for human pose estimation and tracking. In: ECCV, pp 466–481
13. Bertasius G, Feichtenhofer C, Tran D, Shi J, Torresani L (2019) Learning temporal pose estimation from sparsely-labeled videos. In: NeurIPS, pp 3027–3038
14. Sun K, Xiao B, Liu D, Wang J (2019) Deep high-resolution representation learning for human pose estimation. In: IEEE CVPR, pp 5693–5703
15. Ning G, Pei J, Huang H (2020) Lighttrack: a generic framework for online top-down human pose tracking. In: IEEE CVPR workshops, pp 1034–1035
16. Wang M, Tighe J, Modolo D (2020) Combining detection and tracking for human pose estimation in videos. In: IEEE CVPR, pp 11088–11096
17. Cao Z, Simon T, Wei S-E, Sheikh Y (2017) Realtime multi-person 2d pose estimation using part affinity fields. In: IEEE CVPR, pp 7291–7299
18. Raaj Y, Idrees H, Hidalgo G, Sheikh Y (2019) Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In: IEEE CVPR, pp 4620–4628
19. Jin S, Liu W, Ouyang W, Qian C (2019) Multi-person articulated tracking with spatial and temporal embeddings. In: IEEE CVPR, pp 5664–5673
20. Khalid S, Goldenberg M, Grantcharov T, Taati B, Rudzicz F (2020) Evaluation of deep learning models for identifying surgical actions and measuring performance. JAMA Netw Open 3(3):201664–201664
21. Jin A, Yeung S, Jopling J, Krause J, Azagury D, Milstein A, Fei-Fei L (2018) Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: 2018 IEEE WACV, IEEE, pp 691–699
22. Laina I, Rieke N, Rupprecht C, Vizcaíno JP, Eslami A, Tombari F, Navab N (2017)Concurrent segmentation and localization for tracking of surgical instruments. In: MICCAI. Springer, pp 664–672
23. Du X, Kurmann T, Chang P-L, Allan M, Ourselin S, Sznitman R, Kelly JD, Stoyanov D (2018) Articulated multi-instrument 2-d pose estimation using fully convolutional networks. IEEE TMI 37(5):1276–1287
24. Richa R, Balicki M, Meisner E, Sznitman R, Taylor R, Hager G (2011) Visual tracking of surgical tools for proximity detection in retinal surgery. In: international conference on information processing in computer-assisted interventions. Springer, pp 55–66
25. Sznitman R, Richa R, Taylor RH, Jedynak B, Hager GD (2012) Unified detection and tracking of instruments during retinal microsurgery. IEEE PAMI 35(5):1263–1273
26. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the clear mot metrics. EURASIP J Image Video Process 2008:1–10
27. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks. In: international conference on learning representations
28. Simon T, Joo H, Matthews I, Sheikh Y (2017) Hand keypoint detection in single images using multiview bootstrapping. In: IEEE CVPR, pp 1145–1153
29. Santavas N, Kansizoglou I, Bampis L, Karakasis E, Gasteratos A (2020) Attention! a lightweight 2d hand pose estimation approach. IEEE Sens J 21(10):11488–11496
30. 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 ICCV, pp 813–822
31. Zhang J, Jiao J, Chen M, Qu L, Xu X, Yang Q (2017) A hand pose tracking benchmark from stereo matching. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 982–986
32. Gomez-Donoso F, Orts-Escolano S, Cazorla M (2019) Large-scale multiview 3d hand pose dataset. IVC 81:25–33
33. Hadsell R, Chopra S, LeCun Y (2006) Dimensionality reduction by learning an invariant mapping. In: IEEE CVPR. IEEE, vol 2, pp 1735–1742
34. Yan S, Xiong Y, Lin D (2018) Spatial temporal graph convolutional networks for skeleton-based action recognition. In: thirty-second AAAI conference on artificial intelligence
35. Shan D, Geng J, Shu M, Fouhey DF (2020) Understanding human hands in contact at internet scale. In: IEEE CVPR, pp 9869–9878