ShaRPy: Shape Reconstruction and Hand Pose Estimation from RGB-D with Uncertainty

Anonymous ICCV submission

Paper ID 65

Abstract

Despite their potential, markerless hand tracking technologies are not yet applied in practice to the diagnosis or monitoring of the activity in inflammatory musculoskeletal diseases. One reason is that the focus of most methods lies in the reconstruction of coarse, plausible poses, whereas in the clinical context, accurate, interpretable, and reliable results are required. Therefore, we propose ShaRPy, the first RGB-D Shape Reconstruction and hand Pose tracking system, which provides uncertainty estimates of the computed pose, e.g., when a finger is hidden or its estimate is inconsistent with the observations in the input, to guide clinical decision-making. Besides pose, ShaRPy approximates a personalized hand shape, promoting a more realistic and intuitive understanding of its digital twin. Our method requires only a light-weight setup with a single consumer-level RGB-D camera yet it is able to distinguish similar poses with only small joint angle deviations in a metrically accurate space. This is achieved by combining a data-driven dense correspondence predictor with traditional energy minimization. To bridge the gap between interactive visualization and biomedical simulation we leverage a parametric hand model in which we incorporate biomedical constraints and optimize for both, its pose and hand shape. We evaluate ShaRPy on a keypoint detection benchmark and show qualitative results of hand function assessments for activity monitoring of musculoskeletal diseases.

1. Introduction and Related Work

Hand function is affected by musculoskeletal rheumatic diseases. Rheumatoid Arthritis (RA) and Psoriatic Arthritis (PsA) are both common chronic inflammatory diseases, characterized by joint pain and swelling that can result in joint destruction [16]. In view of improved treatment options a more detailed, objective assessment of hand function is desirable, as it can potentially serve as a biomarker for changes in disease activity and patient quality of life [14]. This would allow for early therapy adjustment and potentially improve the prediction of increased risk of joint destruction. In clinical practice, functional assessments are mainly based on subjective questionnaires [23] or manual tests [10] that can discriminate between healthy individuals and patients, but lack sensitivity for disease monitoring over time [22]. The gold standard for objective hand motion assessment is marker-based motion capturing [17]. Other methods use gloves and inertial measurement units [9, 24] to record or monitor hand motion. A major drawback of these technologies is that they are contact-based, time-consuming to set up, and do not provide direct and intuitive visual feedback options. Furthermore, they are not suitable for patient monitoring at home. Hence, simple markerless hand movement assessments based on

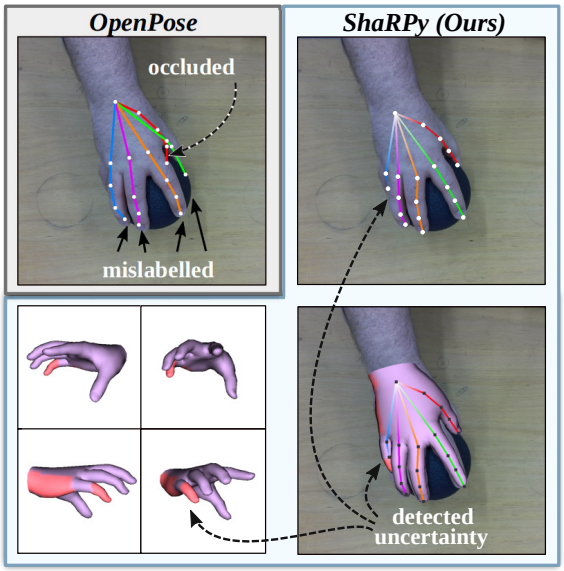

Figure 1: Compared to keypoint approaches, e.g. OpenPose [3, 25], ShaRPy estimates the 3D hand pose and shape, and indicates uncertainty by detecting unobserved and error-prone regions (both visualized in red on the hand surface).

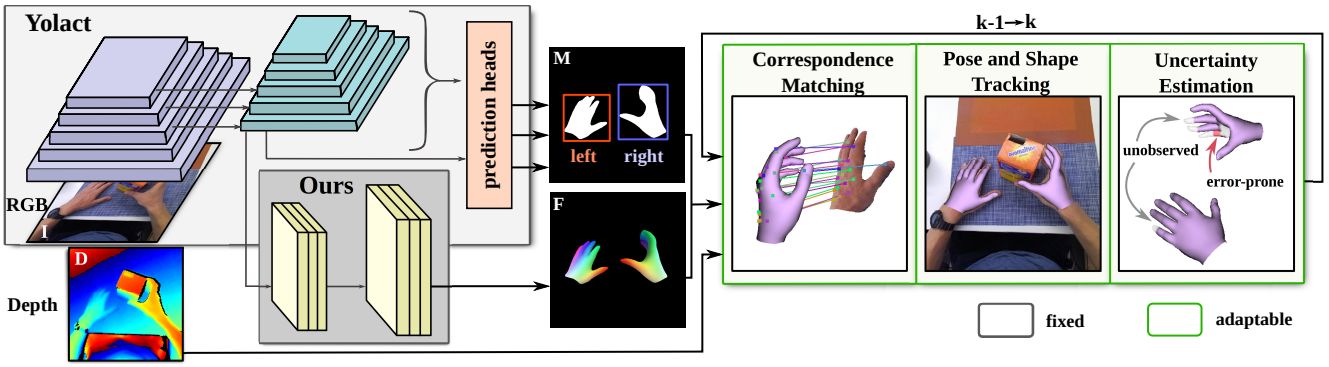

Figure 2: Overview of ShaRPy. First, a network based on Yolact [2] detects hands and regresses features in a correspondence space. The outputs and the depth map are used in a subsequent energy minimization framework for pose and shape estimation. Lastly, we detect uncertainties with respect to the pose parameters.

consumer-friendly sensor systems such as RGB-D cameras are desirable and show promising potential to be applied in the future [20].

In the computer vision community, camera-based hand reconstruction has a rich history [27]. Hand pose estimation algorithms usually reconstruct hands as a set of keypoints [32, 7]. However, the visual interpretability of keypoints is limited (cf. Figure 1) as they do not reflect shape and shape-dependent pose. For example, the neutral posture with all fingers closed of a thick hand is identical to a thin hand with a slight abduction in the Metacarpophalangeal (MCP) joints. Another line of work focuses on estimating the pose of parametric hand models including shape [13, 30, 6, 19]. Commonly, a neural network [13, 30] is trained, which is fixed at inference time and restricted in its generalization capability with respect to unseen shapes, poses, and viewpoints. Alternative approaches are based on energy optimization [6, 19], which can be adjusted to individual video sequences and extended to fit clinical requirements, e.g., including anthropometric hand constraints. Furthermore, in setups with only a single RGB [13, 32, 7] camera, it is challenging to estimate the parameters in a metrically accurate 3D space because the depth of a hand can only be estimated up to a certain scale. To avoid complex setups with multiple cameras, we prefer to use additional depth information of a single RGB-D [6] sensor. The common goal of all the above approaches is to estimate the most plausible pose of the hand and its skeleton. However, in difficult cases (cf. Figure 1), this means that the finger segments can be mislabelled, point into the wrong direction, or are speculated at positions that are not visible. In clinical setups, besides accuracy, it is important to identify and discard unreliable measurements and avoid false positives in the assessment of hand functions.

To tackle all these limitations, we propose, to the best of our knowledge, the first markerless hand tracking method, which provides accurate hand pose *and* shape parameters *and* estimates the uncertainty that remains in those in order to discard unreliable predictions, e.g., when a finger is hidden or its estimate is inconsistent with the observations in the input. Our approach requires only a single RGB-D camera, which makes it easily applicable and allows us to determine a metrically accurate hand shape and pose. ShaRPy makes the following contributions:

- We present the first framework that utilizes dense correspondence predictions to estimate uncertainty through unobserved and error-prone regions of a parametric hand model after shape and pose optimization.
- We introduce a novel correspondence space with semantic encodings, which can be directly transformed into a hand part segmentation. The transformation enables a consistent coarse-to-fine mapping between hand segments and their respective features within each segment, and is utilized for precise correspondence matching and uncertainty estimation.
- We demonstrate the benefits of our approach in the context of markerless hand function assessments as a method to monitor the activity of musculoskeletal rheumatic diseases as well as through a state-of-the-art pose estimation benchmark.

2. Overview

An overview of ShaRPy is shown in Figure 2. First, a pre-trained multi-task network [2] predicts for each hand in an RGB image $\boldsymbol{I}$ its bounding box, a label indicating whether it is the left or right hand, a segmentation mask $\boldsymbol{M}$, and a correspondence image $\boldsymbol{F}$. The correspondence image assigns each pixel of the hand to a unique feature in a novel correspondence space with semantic encodings (Section 3). Subsequently, the optimal pose and shape parameters of a parametric hand model are found in a two-stage energy

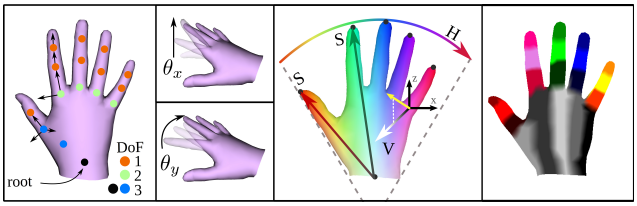

Figure 3: Left: The anatomical MANO model with exemplary movements of joint i in the sagittal (θ_x^i) and coronal plane (θ_y^i). Middle: Dense correspondence encoding. Right: Segmentation sets S_{3d}^i computed from correspondence space.

minimization framework using the additional depth image $\boldsymbol{D}$ (Section 4). Lastly, we estimate the uncertainty through the calculation of unobserved and error-prone regions on the surface of the hand model and visualize the results accordingly (Section 5). Our tracking approach leverages the advantages of video data and reuses the network output, e.g. the Region-Of-Interest (ROI) defined by the bounding boxes, and hand model predictions of the previous frame at timestep $k-1$ to improve the predictions in the current frame k.

Hand Model. We employ the widely adopted MANO model [21] as a parametric representation of the hand. The model is represented by a set of vertices $\mathcal{V} \subseteq \mathbb{R}^3$, deformed by a kinematic tree of 15 finger joints $\mathcal{J} \subseteq \mathbb{R}^3$, and a root wrist joint. The rigid motion of the wrist is described by the translation vector $\mathbf{t} \in \mathbb{R}^3$ and rotation $\mathbf{R} \in \mathbb{R}^3$ in axis-angle notation. Similarly, the per-joint rotations are denoted as the pose $\boldsymbol{\theta} \in \mathbb{R}^{3|\mathcal{J}|}$, and the hand shape is parameterized by $\boldsymbol{\beta} \in \mathbb{R}^{10}$. A linear function maps the pose and shape parameters to joints and, subsequently, to vertices. As the model is not anatomically constrained, the orientation of the joints is not aligned with the natural bone structure. Together with the high number of 3 Degrees-of-Freedom (DoF) per joint, the parametrization can lead to unnatural poses. Inspired by [30], we rephrase the orientation of a per-joint pose such that the respective joint moves within the sagittal, coronal, and transverse plane. Furthermore, we propose to limit the DoF per joint with respect to anatomy considering the special case of the thumb. In total, we reduce the number of optimizable pose parameters from $3 \cdot |\mathcal{J}| = 45$ to 23. The optimized MANO model is shown in Figure 3 and enables an anatomically correct pose parametrization. Please note that, in the following sections, we use $\theta \in \mathbb{R}^{23}$ to denote the *anatomically optimal* pose.

3. Dense Correspondence with Semantic Encodings

Our goal is to fit the MANO model such that it best describes the observations in an RGB-D image. To this end, we establish correspondences between a pixel (x, y) and a vertex $\boldsymbol{v} \in \mathcal{V}$ through a novel, shared canonical correspondence space embedded in $[0,1]^3$. For this, we define the function $\boldsymbol{c}\colon \mathcal{V} \rightarrow [0,1]^3$, which maps $\boldsymbol{v}$ to its coordinate in the correspondence space. As depicted in Figure 3, the space is encoded into a Hue-Saturation-Value (HSV) color cylinder wrapped around the flat rest pose of the model, aligned such that the axes describe semantic features of the hand. The hue describes the angle of a vertex in a circle within the coronal plane and encodes the finger type. We scale the range of $[0°, 360°)$ to lie within the extent of the MANO model to ensure space compactness. This is important to distinguish between different fingers as small differences in values can lead to wrong assignments during the correspondence prediction and matching (see Section 3.2). The saturation is computed on each finger separately and encodes the corresponding finger segment on an axis between the origin of the hand wrist and the fingertip. To distinguish between the front and back of the hand, the value axis encodes the surface normal along the y-axis. In summary, the correspondence space encodes both, spatial and semantic hand features while being compact, continuous, and deterministic to compute. The semantic encoding enables us to define a function $d\colon [0,1]^3 \rightarrow \{1, ..., 20\}$ that computes a discrete segmentation label out of the continuous space, which is later used in Section 3.2 and Section 5. Figure 3 shows the corresponding segmented vertex sets $\mathcal{S}_{3d} = \{S_{3d}^i\}_{i=1}^{20}$ with $S_{3d}^i = \{\boldsymbol{v} \in \mathcal{V} \mid d(\boldsymbol{c}(\boldsymbol{v})) = i\}$, of which 15 refer to the three segments of each finger, and the remaining divide the large area of the wrist into 5 per-finger regions.

3.1. Correspondence Regression

As depth-only datasets are limited in availability and generalization across depth images of different sensor types is challenging, we leverage a variety of RGB(-D) datasets [18, 31, 11, 6, 7] to train our correspondence regression network only on RGB data in a fully supervised manner, and leverage the additional depth component only at test-time during energy minimization. We use a mixture of automatically and semi-automatically labeled ground-truth MANO parameters to transform the models to their position in the image and render the parts of the visible surface to obtain ground-truth correspondence images $\boldsymbol{F}$. In order to detect inconsistent per-pixel predictions of $\boldsymbol{F}$ at inference time and relate them to certain regions of the hand, an additional segmentation map of the visible parts of the hand is required. As our novel correspondence space enables the direct conversion from unique coordinates to coarse hand segments, it is not necessary to predict an additional segmentation mask, which could potentially lead to inconsistent per-pixel predictions with $\boldsymbol{F}$ otherwise. Instead, for each hand visible in an image $\boldsymbol{I}$, our framework only pre-

dicts dense correspondences, of which we compute a segmentation set $S^i_{2d} = \{(x, y) \mid d(\boldsymbol{F}(x, y)) = i\}$ of pixels (x, y) for each segmentation label i.

Our regression network is an extension of Yolact [2] with an additional branch for correspondence prediction, which is trained by minimizing the smooth L1 loss between the predicted and ground-truth correspondence value of each pixel within the ground-truth segmentation mask of the hand. At inference time, we multiply the correspondence values with the predicted mask $\boldsymbol{M}$ to acquire per-pixel correspondences only for the hand.

3.2. Correspondence Matching

Correspondence pairs are established by comparing each predicted $\boldsymbol{c}_p = \boldsymbol{F}(x, y)$ at pixel (x, y) with $\boldsymbol{c}_v = \boldsymbol{c}(\boldsymbol{v})$ of every MANO vertex $\boldsymbol{v}$. A common method to find a match is a traditional nearest-neighbor search [19]. In particular for hands, this method can result in wrong correspondence pairs at positions in between fingers. This is because, contrary to the continuous nature of the correspondence space, the assignment of a pixel to a vertex of a specific finger is a discrete problem. We improve the quality of correspondence pairs by using both, the correspondence space and its discrete segmentation, and compute nearest-neighbor matches only within the sets of segmented vertices $\mathcal{S}^i_{3d}$ and segmented pixels S^i_{2d} that share the same segmentation label. In other words, we first reject possible matches on the coarse segmentation level in case they do not share the same label and, subsequently, compute matches in the correspondence space. A match between $\boldsymbol{c}_p$ and $\boldsymbol{c}_v$ is used to construct a pair $(\boldsymbol{p}, \boldsymbol{v})$ of 3D correspondences between $\boldsymbol{v}$ and an image point $\boldsymbol{p} \in \mathbb{R}^3$, computed from the back-projection of the depth value at $\boldsymbol{D}(x, y)$. Since $\boldsymbol{c}_p$ is predicted in the view of the RGB camera, it is not exactly aligned with the pixel positions of $\boldsymbol{D}$. Particularly at the edges of the hand silhouette, the predictions can map to erroneous points of the background. Hence, we first discard pairs, in which $\boldsymbol{D}(x, y)$ deviates too far from the median depth of the hand, determined by a threshold t_d. Second, we filter out points at silhouette edges with degraded and noisy depth by inspecting whether the angle of the point-wise normal computed from $\boldsymbol{D}$ exceeds a given threshold t_n. Lastly, we discard all pairs $(\boldsymbol{p}, \boldsymbol{v})$, of which the Euclidean norm of their difference exceeds the 3D distance threshold t_{3d}. The final 3D correspondence set is denoted by $\mathcal{C}_{3d}$.

4. Pose and Shape Tracking

In this stage, we solve an energy-minimization problem to obtain the optimal MANO parameter set $\Omega^k = (R^k, t^k, \theta^k, \beta^k)$ at timestep k:

$$\arg\min_{\Omega^k} \left[\omega_{3d}\lambda E_{3d}(\mathcal{C}_{3d}) + \omega_{2d}E_{2d}(\mathcal{C}_{2d}) + E_{reg}(\Omega^k, \Omega^{k-1})\right]$$

We denote the respective weights of a term E_* as ω_* and define $\lambda = \exp(J + 1)$, where J is the Jaccard index of the predicted mask $\boldsymbol{M}$ and the mask $\boldsymbol{M}_v$ of the rasterized MANO model. We generate $\boldsymbol{M}_v$ by using the differentiable rasterizer Nvdiffrast [12]. E_{3d} and E_{reg} are similar to Mueller et al. [19]: The data term E_{3d} consists of a point-to-point and point-to-plane error. The regularization term E_{reg} enforces plausible poses and shapes, as well as temporal smoothness, and consists of E_{shape}, E_{pose}, and E_{temp}. In contrast to [19], we use the anatomically rephrased orientations of the MANO model such that E_{pose} enforces poses within anatomical limits. Furthermore, we introduce the term E_{2d} defined on the set of valid pixels $\mathcal{C}_{2d}$ within $\boldsymbol{M}$ and $\boldsymbol{M}_v$. For each pixel $(x, y) \in C_{2d}$, the term penalizes the squared L2 norm between $\boldsymbol{F}(x, y)$ and $\boldsymbol{F}_v(x, y)$, where F_v is the correspondence image of the rasterized hand. In other words, E_{2d} enforces the MANO model to lie within the predicted hand silhouette and provides a more accurate estimation of $\boldsymbol{\beta}$ compared to E_{3d}. In our energy minimization framework, we distinguish between the *Initialization* phase, which is only executed in the first frame or when the tracking is lost, and the *Refinement* phase, in which we iteratively minimize E. During initialization, we first solve the orthogonal Procrustes problem to obtain the initial wrist parameters $\boldsymbol{R}$ and $\boldsymbol{t}$. Secondly, we make use of an implicit pose prior to initialize $\boldsymbol{\theta}$ with plausible parameters. For this purpose, we transform the anatomically rephrased $\boldsymbol{\theta}$ into a PCA space, which we pre-compute from annotated RGB(-D) datasets [18, 31, 11, 6, 7]. Then, we solve the energy formulation with respect to the PCA pose parameters in order to obtain a plausible initialization of $\boldsymbol{\theta}$. As the PCA pose space is not expressive enough to capture the high variance of different hand poses, we refine $\boldsymbol{\theta}$ in the subsequent Refinement stage.

5. Uncertainty Estimation

As mentioned in Section 1, the generalization capability of data-driven pose and shape estimation approaches is limited with respect to inputs that do not lie within the learned data distribution, e.g., unseen hand poses or viewpoints. Our approach poses no exception to this general limitation and we observe correspondence mispredictions that exhibit inconsistencies in the anatomic structure of the hand, which is encoded by the correspondence space. These inconsistencies are not only noticeable visually (see Figure 4) but also during energy minimization. More specifically, we experience high residuals in regions, where it is not possible to optimize the parameters of the anatomically constrained MANO model such that its surface is optimal with respect to the position in the image given by the pixels of the cor-

Figure 4: Correspondence predictions (P) on images from H2O [11] compared with their ground-truth (GT). Our network was trained on HO3D [6], InterHand2.6M [18], H_2O-3D [7] and FreiHAND [31]. Inconsistencies in the regressed coordinates are highlighted in red.

respondence pairs. Correspondence coordinates with a significant deviation from their actual position in the space are assigned to a wrong segmentation label through the discretization of $d(\cdot)$. Hence, hand segments can either be over-saturated with mispredicted correspondences or have no correspondences at all despite being visible in the input image, as depicted in Figure 4. Based on these observations, we obtain an uncertainty value u_i for each segment i on the surface of the MANO model given by the segmentation sets S^i_{3d}, which are computed from the predicted correspondence image $\boldsymbol{F}$. We compute the uncertainty value such that:

$$u_i = \begin{cases} 1 & \text{if segment } i \text{ unobserved or error-prone} \\ 0 & \text{else} \end{cases}$$

Since a segment relates to the set of vertices deformed by a particular joint, we can directly infer uncertainty with respect to its respective pose parameter. We consider a segment i as unobserved if:

$$\frac{|\mathcal{V}^i_{vis}|}{|S^i_{3d}|} < \tau_v, \quad \text{with} \quad \mathcal{V}^i_{vis} = \{v \in \mathcal{S}^i_{3d} \mid (*, v) \in \mathcal{C}_{3d}\}$$

Further, we consider a segment i as error-prone if:

$$\frac{|\mathcal{P}_{2d}|}{|S^i_{2d}|} > \tau_{2d} \quad \text{or} \quad \frac{|\mathcal{P}_{3d}|}{|S^i_{3d}|} > \tau_{3d}$$

We define $\mathcal{P}^i_{2d} = \{(x, y) \in \mathcal{S}^i_{2d} \mid (x, y) \in \mathcal{C}_{2d} \wedge E_{2d}(x, y) > \varepsilon_{2d}\}$ as the set of error-prone pixels and, analogously, $\mathcal{P}^i_{3d} = \{\boldsymbol{v} \in S^i_{3d} \mid (*, \boldsymbol{v}) \in C_{3d} \wedge E_z(\boldsymbol{v}) > \varepsilon_{3d}\}$ as the set of error-prone vertices. The term $E_z(\boldsymbol{v})$ is defined as the average L1 loss between the z-axis values of all pairs in $\mathcal{C}_{3d}$, in which $\boldsymbol{v}$ is included.

6. Results

Our network is implemented and trained in PyTorch. For the evaluation, we apply three different training procedures, denoted as *V1*, *V2*, and *V3*. In V1, we exclusively train on the H2O [11] dataset. In V2, we train on all previously mentioned datasets [11, 6, 7, 18, 31]. In V3, we exclude H2O and train on the remaining data [6, 7, 18, 31]. At inference time, the network is embedded together with the rest of the pipeline into a common C++ framework. During shape and pose estimation, we initialize the tracking by using libTorch's L-BFGS optimizer and then iteratively refine the energy with Adam. The results are divided into two experiments. First, we show the clinical applicability of our setup (with V2) on a male, 61 years old PsA patient (Disease Activity in Psoriatic Arthritis score: 17.52) and demonstrate the reliability to discard invalid pose predictions through the detection of uncertainty. Second, we quantitatively (V1 and V3) and qualitatively (V2) compare the accuracy of our method with the state-of-the-art (SOTA).

6.1. Clinical applicability

Similar to clinical practice, we recorded a sequence of the finger adduction and abduction together with the finger hyperextension and assess the hand function by measuring the angles of the fingers, using the middle finger as a reference. We achieve this by projecting the segments between the proximal interphalangeal joints (PIP) and MCP joints onto the wrist plane and computing the angle deviation from the PIP-MCP segment of the middle finger. The results are depicted in Figure 5. We are further able to visualize the finger hyperextension due to the depth information, which is not possible in RGB-only approaches. Next, we recorded the patient holding a ball and rotating the wrist around the camera. As we can assume that the fingers hardly move during this task, we expect corresponding results in the finger angles. We plot the angles of the pose θ around the MCP of the thumb and the index and filter out all measurements, in which one of the respective finger segments is marked as uncertain within three consecutive frames. We compare the results with unfiltered angle measurements and perceive a significant decrease in angle variance from $112.55°$ to $18.16°$ on the middle finger and from $125.29°$ to $37.84°$ on the thumb, which was less visible and mainly close to silhouette edges in the depth map.

6.2. Comparison with State-of-the-art

Since there is no established evaluation method for *dense* pose and *shape* estimation with *uncertainty* estimation in *clinical* applications, we compare our method with the SOTA on pose estimation. Therefore, we evaluate the accuracy of ShaRPy on the H2O [11] dataset, which contains egocentric RGB-D sequences of healthy subjects most visually close to a clinical setting. In Figure 5, we compare the qualitative results of V2 with OpenPose [3, 25]. For a quantitative comparison, our results are submitted and objectively evaluated on a public leaderboard. The benchmark

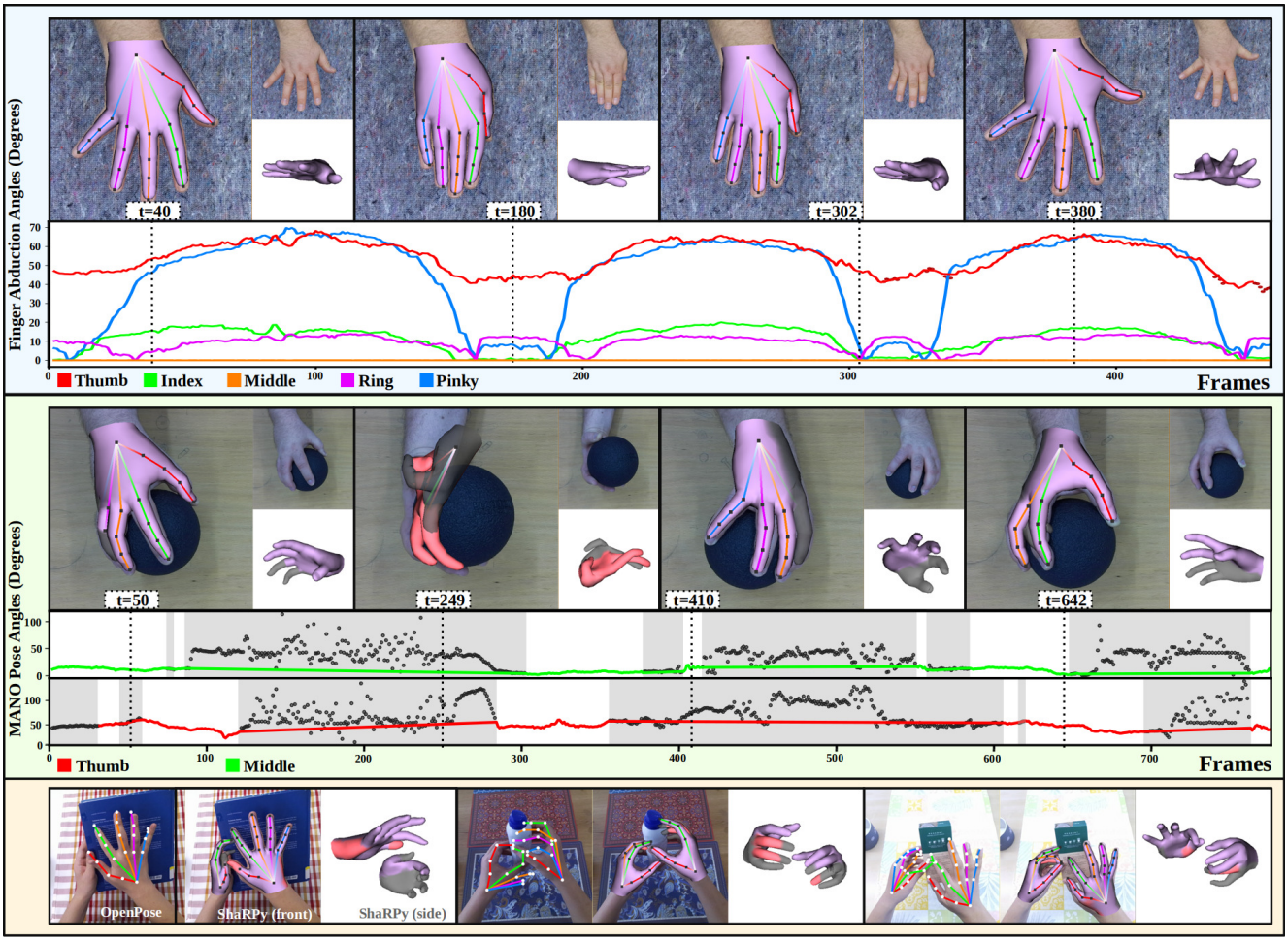

Figure 5: Top: Images and angle results from an abduction and adduction sequence (repeated 3×). Middle: Results of the rotating ball sequence. Unobserved (grey hand surface) or error-prone (red surface) poses are listed as disconnected grey dots in the plots. Bottom: Comparison of ShaRPy with OpenPose [3, 25].

is tailored to RGB keypoint-based methods and evaluates the plausibility of poses in the presence of strong occlusions. Table 1 summarizes the results with respect to the Mean End-point Error (MEPE) and the Percentage of Correct Keypoints (PCK). In summary, ShaRPy places first or third on the leaderboard, even though we did not design our system specifically for a keypoint-based pose estimation challenge, do not focus on plausibility, and, solve a more challenging problem of indirectly estimating the pose through shape along with the shape itself. On top of that, we show the generalization ability of our version V3, which outperforms most methods by placing third.

7. Conclusion

In this work, we proposed the first markerless hand tracking approach, which calculates uncertainty in the pose estimates. Our approach combines a data-driven dense correspondence predictor with a flexible, generative energy minimization framework to estimate the optimal hand pose and shape that best explains the given observations. Further, we detect uncertain poses through the detection of unobserved and error-prone surface segments. We demonstrate through quantitative and qualitative results that our approach provides outstanding pose estimation accuracy, on top of its generalization to both, unknown datasets of healthy individuals and patient data. Furthermore, we provide results of clinical hand function assessments and show that, compared to other markerless approaches, our approach has no limitation in terms of its applicability and, instead, includes more favorable properties such as additional shape estimation and the robust filtering of uncertain poses. We believe our approach can be used to drive further research in the context of markerless tracking in clinical applications.

	MEPE (mm)↓		3D PCK@15mm↑		3D PCK@30mm↑	
	left	right	left	right	left	right
Hasson et al. [8]	39.56	41.87	-	-	-	-
Tekin et al. [26]	41.32	38.86	-	-	-	-
Kwon et al. [11]	41.45	37.21	-	-	-	-
Aboukhandra et al. [1]	36.80	36.50	-	-	-	-
Cho et al. [4]	24.40	25.80	-	-	-	-
Wen et al. [28]*, [29]	35.02	35.63	12.67	2.98	43.71	37.12
Cho et al. [5]*	14.40	15.90	70.75	54.61	93.81	95.08
Luo et al. [15]*	20.80	24.70	40.77	32.29	80.36	73.56
Ours (V1)	**20.47**	**19.07**	**21.04**	**27.81**	**92.81**	**94.73**
Ours (V3)	28.62	28.42	12.95	16.64	81.61	86.15

Table 1: Results on the H2O [11] hand pose challenge. For each metric, we indicate whether higher results (↑) or lower results (↓) are better. The best results among accepted conference publications are highlighted in bold. For completeness, we also list workshop contributions, which are tailored towards the H2O challenge, and denote them with *.

.

References

[1] Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1001–1010, 2023.

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, 2019.

[3] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[4] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4769–4778, June 2023.

[5] Hoseong Cho, Donguk Kim, Chanwoo Kim, Seongyeong Lee, and Seungryul Baek. Transformer-based global 3d hand pose estimation in two hands manipulating objects scenarios. *arXiv e-prints*, page arXiv:2210.11384, Oct. 2022.

[6] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[7] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11080–11090, 2022.

[8] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. *CoRR*, abs/2004.13449, 2020.

[9] J Henderson, J Condell, J Connolly, D Kelly, and K Curran. Review of Wearable Sensor-Based Health Monitoring Glove Devices for Rheumatoid Arthritis. *Sensors (Basel)*, 21(5), 2021.

[10] S C Higgins, J Adams, and R Hughes. Measuring hand grip strength in rheumatoid arthritis. *Rheumatol International*, 38(5):707–714, 2018.

[11] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[12] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020.

[13] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021.

[14] A M Liphardt, E Manger, S Liehr, L Bieniek, A Kleyer, D Simon, K Taskilar, M Sticherling, J Rech, G Schett, and A J Hueber. Similar Impact of Psoriatic Arthritis and Rheumatoid Arthritis on Objective and Subjective Parameters of Hand Function. *ACR Open Rheumatology*, 2(12):734–740, 2020.

[15] Weixin Luo, Shuqiang Cao, Bairui Wang, Wei Zhang, Xiaolin Wei, and Lin Ma. Yolov7-3d: One-stage monocular 3d hand pose estimation. In *2022 IEEE International Conference on Computer Vision (ICCV) Workshops: Human Body, Hands, and Activities from Egocentric and Multi-view Cameras (HBHA)*, 2022.

[16] Joseph F Merola, Espinoza Luis R, and Fleischmann Roy. Distinguishing rheumatoid arthritis from psoriatic arthritis. *RMD Open*, 2018.

[17] C D Metcalf, S V Notley, P H Chappell, J H Burridge, and V T Yule. Validation and application of a computational model for wrist and hand movements using surface markers. *IEEE Trans Biomed Eng*, 55(3):1199–1210, 2008.

[18] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline

for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020.

[19] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics (TOG)*, 38(4), 2019.

[20] Uday Phutane, Anna-Maria Liphardt, Johanna Bräunig, Johann Penner, Michael Klebl, Koray Tascilar, Martin Vossiek, Arnd Kleyer, Georg Schett, and Sigrid Leyendecker. Evaluation of Optical and Radar Based Motion Capturing Technologies for Characterizing Hand Movement in Rheumatoid Arthritis-A Pilot Study. *Sensors (Basel)*, 21(4), 2021.

[21] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.

[22] M Rydholm, I Wikström, S Hagel, L T H Jacobsson, and C Turesson. The Relation Between Disease Activity, Patient-Reported Outcomes, and Grip Force Over Time in Early Rheumatoid Arthritis. *ACR Open Rheumatology*, 1(8):507–515, 2019.

[23] F Salaffi, M Di Carlo, S Farah, D Marotto, F Atzeni, and P Sarzi-Puttini. Rheumatoid Arthritis disease activity assessment in routine care: performance of the most widely used composite disease activity indices and patient-reported outcome measures. *ACR Open Rheumatology*, 92(4), 2021.

[24] C Salchow-Hömmen, L Callies, D Laidig, M Valtin, T Schauer, and T Seel. A Tangible Solution for Hand Motion Tracking in Clinical Applications. *Sensors (Basel)*, 19(1):1199–1210, 2019.

[25] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[26] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4506–4515, 2019.

[27] Edith Tretschk, Navami Kairanda, Mallikarjun B R, Rishabh Dabral, Adam Kortylewski, Bernhard Egger, Marc Habermann, Pascal Fua, Christian Theobalt, and Vladislav Golyanik. State of the art in dense monocular non-rigid 3d reconstruction, 2022.

[28] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recognition from Egocentric RGB Videos. *arXiv e-prints*, page arXiv:2209.09484, Sept. 2022.

[29] Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, and Wenping Wang. Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21243–21253, June 2023.

[30] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086, 2021.

[31] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In Christian Bauckhage, Juergen Gall, and Alexander Schwing, editors, *Pattern Recognition*, pages 250–264, Cham, 2021. Springer International Publishing.

[32] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017.