Dynamic Multiview Refinement of 3D Hand Datasets using Differentiable Ray Tracing

Giorgos Karvounas[1,2] Nikolaos Kyriazis[1] Iason Oikonomidis[1] Antonis Argyros[1,2]

[1]Institute of Computer Science, FORTH, Greece
[2]Computer Science Department, University of Crete, Heraklion, Greece
{gkarv, kyriazis, oikonom, argyros}@ics.forth.gr

Abstract

With the increase of AI applications in the field of 3D estimation of hand state, the quality of the datasets used for training the relevant models is of utmost importance. Especially in the case of datasets consisting of real-world images, the quality of annotations, i.e., how accurately the provided ground truth reflects the true state of the scene, can greatly affect the performance of downstream applications. In this work, we propose a methodology with significant impact on improving ubiquitous 3D hand geometry datasets that contain real images with imperfect annotations. Our approach leverages multi-view imagery, temporal consistency, and a disentangled representation of hand shape, texture, and environment lighting. This allows to refine the hand geometry of existing datasets and also paves the way for texture extraction. Extensive experiments on synthetic and real-world data show that our method outperforms the current state of the art, resulting in more accurate and visually pleasing reconstructions of hand gestures.

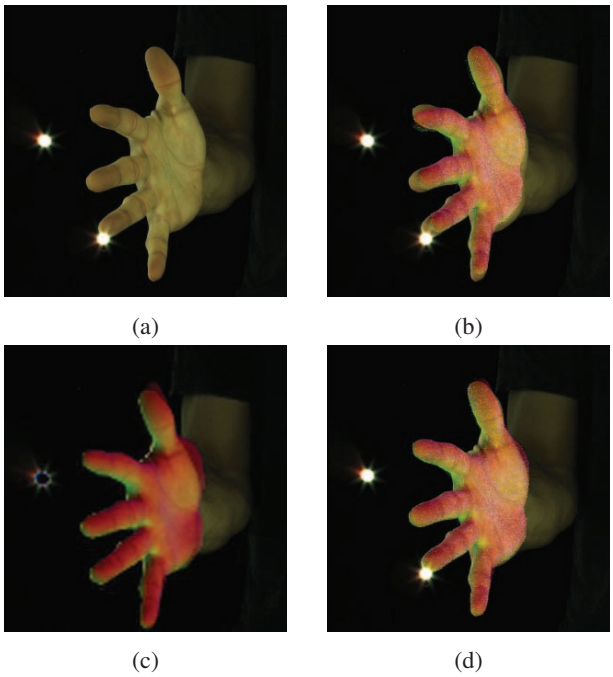


Figure 1: Using (a) multi-view image sequences (videos) of a moving hand, our method (DMVR) can refine (b) an initial inaccurate hand pose and geometry estimate to (d) an estimate of higher fidelity, beyond (c) what was attainable with the SOTA approach in [24] (SMVR). Through DMVR, datasets such as InterHand2.6M can be greatly improved.

1. Introduction

The estimation of hand shape, pose, and appearance from visual input has been a topic of great interest in the fields of computer vision, Human Computer Interaction and robotics for many years, due to the significant impact of a potential solution to a wide range of applications. The problem presents several challenges including handling of ambiguities, occlusions, variability of hand shape, movement and context, and the requirement for real-time performance.

The Visual AI revolution that started in 2012 [27] also revolutionized the field of 3D hand state estimation through deep learning. This approach gradually gained popularity since 2014 [68]. Soon, the research community actively acknowledged the value behind collecting and training on large amounts of high-quality annotated data, which has led to the proliferation of relevant datasets (see Section 2.2). Still, the task of establishing new datasets is always relevant, as there is no single dataset to serve all purposes. Additionally, improvements in methodologies call for better annotation quality, constantly feeding a need for new, better annotated datasets, essentially rendering current datasets of inadequate quality. This becomes immediately apparent when looking at State-of-the-Art (SOTA) datasets, such as InterHand2.6M. In fact, both in [40] and in [24], one can see

(a) LISA [9]

(b) Our method

Figure 2: We are motivated by how LISA [9] aspires to do monocular estimation of geometry and color. It is a very challenging task, which is reflected in the superimposition of their results. For the same pose, DMVR can provide ground truth that is of significantly better quality than the one LISA was trained on, improving their results.

how easily discernible the culprit is, through visual inspection. Utilizing such datasets inherently requires overcoming the shortcomings in annotation quality, as exemplified in Fig. 2. Simultaneously, as involvement with AI intensified, the "data rules" campaign emerged, championed by Andrew Ng [43], Anandkumar [1], Gil Press [52], Michael I. Jordan [22] and others. Along these lines, Karvounas *et al*. [24] set out to address the issue at its source, trying to "fix" the InterHand2.6M dataset and, thus, relieving all further research from annotation issues. They introduced high-fidelity ray tracing to extract as much detail as possible for realistic differentiable rendering, in an effort to maximize the accuracy and fidelity of the resulting annotations.

This work shares this motivation and aspires to further improve datasets such as InterHand2.6M (see Fig. 1). InterHand2.6M, as well as most contemporary hand-related datasets, employ MANO [59] as a highly regularized yet expressive 3D hand representation. This is attractive for several reasons, so, better versions of datasets, *i.e*. more accurate MANO representations, are of significant value. The work in [24] brought together converging evidence from multiple views, through ray tracing, to super-sample and constrain even more the process of fine-tuning inaccurate 3D annotations in an image-based fashion. We take this notion several steps further by (a) extending the constraint extraction across time, and, (b) by properly accommodating different types of constraint contributions through departing from simple diffusion reflectance and going towards full-blown radiance. In turn, this clarifies the contributions of lighting and hand texture in the formation of image sequences of moving hands. We consider this a necessary feature, as otherwise, different observations of the same parts of hand texture would clash, due to the effect of light, *i.e*. changes in shadowing and shading as a lit hand moves. The end result is that, as far as SOTA is concerned, fewer cameras are now needed to deliver significantly better results, yielding a version of InterHand2.6M that is better than before and has the prospect of still becoming even better in the near future. A noteworthy side effect is that hand textures where light has been factored-out can now be considered, which opens entirely new improvement paths. Put plainly, one might be better off observing a moving hand with fewer cameras, getting not only an accurate 3D estimation, but also more complete (because, as the hand moves, more of its surface is revealed) and lighting-free hand textures.

2. Related Work

3D Hand pose estimation from markerless visual input is a significant and challenging problem with a wide range of applications. Researchers have been exploring it for decades [55], and it continues to be an active area of research today [37, 20, 17, 62, 61, 74, 16, 57, 45].

Due to the complexity of the problem, early approaches for hand pose estimation primarily relied on multiview input to mitigate ambiguities caused by occlusions. This approach has regained popularity in recent years [55, 11, 46, 48, 69, 18, 17, 60, 72], along with stereo input [49, 56, 32]. With the widespread availability of commodity depth sensors around 2010, research efforts also focused on monocular depth or RGBD input [47, 26, 63, 53, 65, 35, 64, 44, 70, 37, 14, 57]. However, the emergence of deep learning shortly thereafter, coupled with the availability of hand-related datasets, led to the development of robust approaches that can effectively estimate 3D hand pose from monocular RGB input [58, 75, 50, 5, 41, 20, 15, 13, 66, 28, 4, 74, 16, 61, 62, 51, 6, 36].

The most relevant work to the present one is [24]. Our work builds upon [24] to incur two types of significant contributions, namely, the dramatic increase in integration level (from images to videos), as well as the quality of the end results, *i.e*. the amount of improvement on given input. To better serve the comparison between the two, we will refer to [24] as Static Multi-View Refinement (SMVR), and to our work as Dynamic Multi-View Refinement (DMVR).

2.1. 3D Hand Reconstruction

De la Gorce et al. [10] proposed the first holistic method to reconstruct a hand from a scene using a monocular RGB camera. Due to the fact that differentiable renderers were not available at that time, the authors had to create custom implementations of many components of the rendering pipeline in order to accommodate the optimization task. More recently, using a differentiable render [25] and MANO [59] with the combination of iterative refinement, Baet et al. [2] proposed a deep learning-based method to estimate the pose and shape of an observed hand from

RGB images. Zhang et al. [73] proposed an end-to-end refinement-based framework for recovering the shape of the hand. Using synthetically textured hands, Boukhayma et al. [3] proposed a DNN to recover the mesh of a hand from an RGB image in the wild. Using a higher resolution hand model than MANO, Kulon et al. [29] created the first graph morphable model of the human hand. Lv et al. [34] proposed HandTailor, a lightweight CNN based 3D hand mesh recovery approach. The first self-supervised 3D hand reconstruction pipeline was proposed by Chen et al. [8]. The input to their proposed network was an RGB image of a hand. Using encoders, the method estimated the texture of the hand, the light of the scene and the parameters of the MANO model along with the camera. SeqHAND [71], estimated the 3D hand pose and shape, exploiting temporal information directly from sequential RGB images, considering the appearance of a hand and incorporating motion information. Chen et al. [7] proposed a hand model with 12,337 vertices and a Deep Learning architecture, implicitly learning the texture and the reflectance of the hand, using a self-occlusion-aware shading field, without taking account the lighting conditions of the scene. Using a NeRF architecture, Corona et al. [9] reconstructed the shape and the appearance of a hand using temporal information. However, the network cannot capture the details of the appearance and resolve some depth ambiguities, despite using multiview. Finally, HARP [23] used a monocular video of a masked hand and they approximated the self-shadows of the hand using a rasterizer. Among these works, those that require graphics rendering do not use ray tracing as we do, but resort to simpler approaches such as rasterization or implicitly learned appearance synthesis. Furthermore, most of these works rely on datasets, and therefore on the quality of their annotations. We advocate that there is room for better results if these approaches utilize better annotated datasets.

2.2. Datasets

Mueller et al. [42] proposed two datasets, the SynthHands and the EgoDexter, a synthetic and an egocentric respectively. ObMan is a synthetically generated hand-object interaction dataset proposed by Hasson et al. [19] with 141k training samples, 6.4k validation samples and 6.2k testing samples. FreiHand, the first multi-view hand-object dataset with 3D hand pose and shape annotations was proposed by Zimmermann et al. [76]. The dataset contains recordings of 32 subjects performing gestures or interactions with objects, using 8 RGB cameras. Kulon et al. [28] created YouTubeHands, a dataset that has been automatically annotated, sourced from YouTube. It consists of 54,000 images for training and 1,500 images each for testing and validation purposes. InterHand2.6M is the largest so far RGB 3D hand pose estimation dataset proposed by Moon et al. [40]. The dataset is captured using 80 to 140 high resolution calibrated cameras from 28 subjects performing a variety of poses with one or both hands. For the annotation, the authors employ a semi-automatic approach, which is a combination of manual and automatic annotation [38]. The dataset also provides 3D mesh annotations built using MANO [59]. Finally, ARCTIC [12] is the first dataset depicting free-form interactions between hand and articulated objects, with 1.2M images.

2.3. Parametric Hand Models

Parametric models, pertaining to the geometry and the appearance of the hand, are invaluable in the 3D reconstruction of hands. MANO [59] is the most commonly used geometric model, disentangling shape and pose. DeepHandMesh [39] is similar in nature, with the distinction of relying on a more complex and expressive model (neural instead of PCA). Qian et al. [54] propose the Hand Texture Model (HTML) to model hand appearance. The HTML is a statistical model created based on scans of 51 subjects from various ethnicities. For compatibility, the scans are registered to MANO [59]. Piano [31] integrates anatomical components into its model, focusing specifically on the bone structure and corresponding joints. This approach enables it to generate believable hand motions that respect the physical constraints of human anatomy. In a similar vein, Nimble by Li et al. [33] uses MRI scans to inform its hand model. This allows to capture nuanced information about the muscles and bones of the hand. Moreover, Nimble incorporates texture representation of the hand, proposing a comprehensive approach to hand modeling.

3. Methodology

We present our approach for improving the ground truth annotations of a hand pose dataset, such as InterHand2.6M. Accurate annotations of hand poses are critical for training and evaluating computer vision models for hand tracking and pose estimation. However, bundled annotations are often inaccurate and inconsistent, which can hinder downstream applications. To address this issue, we propose a method (Fig. 3) that uses a fine-tuning optimization process to improve the quality and accuracy of the ground truth annotations by tapping the relevant imagery. Specifically, we aim to generate improved annotations that better capture the nuances and variations of hand poses in the dataset. Our objective is to enhance the alignment of 3D annotations across all views, through 2D backprojections, which invariably improves the fidelity of the 3D annotations, themselves.

3.1. Input

We largely model the input to our method after InterHand2.6M [2]. InterHand2.6M comprises several short multi-camera videos, each briefly demonstrating a hand motion or gesture. The input to our method amounts to

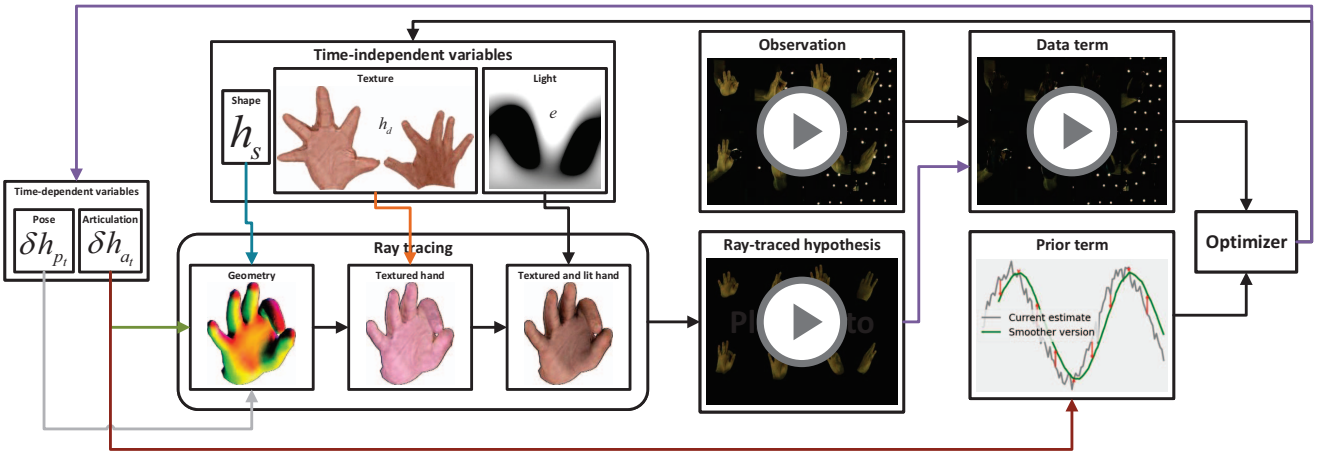


Figure 3: During optimization, a time-independent texture and environment map serve as slack variables for photoconsistency, to accommodate consensus across views. For each multi-view, we estimate the hand geometry as a function of time-dependent pose and time-independent shape, since the hand identity remains the same across the video. We employ ray tracing to produce multi-views, which we compare with observations. The optimizer updates all relevant parameters using their difference.

any of these (or similar) videos. Any such video $V = \{I_{c,t}, c = 1, \ldots, \mathbf{C}, t = 1, \ldots, \mathbf{T}\}$ comprises $\mathbf{C}$ views sampled across $\mathbf{T}$ time instants, for a total of $\mathbf{I} = \mathbf{C} \times \mathbf{T}$ images $I_{c,t}$. Each view c is intrinsically and extrinsically calibrated, and all cameras share a common frame of reference. Within this frame, each time instant t is associated with a 3D annotation h_t in the form of MANO poses, explicitly or implicitly (*i.e.* a MANO pose is not provided but can be reconstructed from other 3D annotations). We are not concerned with other types of annotations.

3.2. Output

Our method yields time-dependent MANO pose and articulation correctives δh_{p_t} and δh_{a_t}, respectively. It also yields a time-independent MANO shape h_s, a hand texture map h_d and an environment map e. The correctives, applied through simple addition on h_t, along with the newly computed shape, hand texture and environment map comprise an improved annotation of the input video V. Treating notation loosely, we refer to the improved result as $\hat{V}$ although it is not a video itself, but rather video annotations. We do this because this information can be used to synthesize images that optimally match the input videos, which is in fact the process we follow to optimize $\hat{V}$.

3.3. Processing

To improve the provided annotations our method solves an image-based fine-tuning optimization problem, across space and time, simultaneously. The objective of this optimization problem is defined as:

$$L(v) = \lambda_{data} L_{data}(v) + \lambda_{prior} L_{prior}(v), \quad (1)$$

and

$$\hat{V} = \arg\min_{v} L(v), \quad (2)$$

where v captures the whole state to be optimized, namely δh_{p_t}, δh_{a_t}, h_s, h_d and e.

3.4. Data term

At any given iteration within an optimization loop, the data term L_{data} quantifies the image-based difference between the current state v and the imagery V. To convert 3D information into comparable images, Monte Carlo Ray Tracing $R_{MC}(\cdot)$ is employed [30] over procedural 3D hand meshes generated through the MANO decoder $D_{MANO}(\cdot)$ [59]. Canny edge detection is employed to put extra weight on the comparison between the image-based edges of the model and the corresponding edges in the imagery. This is relevant as it is not uncommon for hand parts to stand out as intensity discontinuities. This yields the following formulation:

$$L_{data}(v) = \sum_{c=1}^{C} \sum_{t=1}^{T} \| R_{MC}(D_{MANO}(v_{c,t})) - I_{c,t} \|. \quad (3)$$

This basic formulation is augmented, through addition, with the edge enhancement and, through multiplication, by the cross-camera tone mapping methodology also described in [24] (*i.e.*, photo-consistency across views is assumed up to some linear color transform).

A feature that sets apart DMVR and SMVR [24] is the difference in the employed ray tracing mode. In SMVR the rendering channel[1] is set to "diffuse reflectance", which per-

[1]This is a reference to the python-based implementation of Redner.

tains to the fraction of light that is scattered in all directions when it hits a surface. In DMVR the rendering channel is set to "radiance", which pertains to the amount of light energy per unit area per unit solid angle that is emitted or reflected by a surface. The difference between them is that radiance includes both diffuse and specular components, as well as other effects such as shadows and inter-reflections, while diffuse reflectance only includes the diffuse component. Without employing radiance computations one cannot gain access to information that can factorize image formation, *i.e*. light-texture disentanglement, which is critical for the success of DMVR. This sets apart the two in terms of factorization, realism, and, compute, with the more realistic variant (radiance) being almost $3\times$ as costly.

Other differences aside (*e.g*. different terms), an important distinction is that, as opposed to SMVR, in DMVR the summation runs across time, as well. By including all accessible constraints and by distinguishing between time-dependent and time-independent variables, we accommodate all observational cues in a cross-pollinating fashion.

3.4.1 Hand Texture Map

As in [24], we employ a hand texture slack variable to accommodate the photo-consistency assumption at the surface of the hand. However, it's important to note that [24] could not work across time unless the hand texture is decoupled from lighting. Otherwise, the photo-consistency assumption that premises the formulation would be violated.

Similarly to [24], we adopt a multi-Level of Detail (LoD) formulation. Every texel amounts to a blend of increasingly larger 2D images that are interpolated to match a predetermined texture size. This formulation allows the scheduling of LoD during optimization so that rough information is first extracted, and then more detail is accommodated.

We use Redner [30] for rendering the hand, which not only allows gradients to be routed to texture updates in a disentangled fashion, but also ensures that they are correct in terms of sampling. Aliasing can be an issue when sampling textures. The use of Redner mitigates this problem and improves the quality of the rendered images.

3.4.2 Environment Map

As already mentioned, a key feature of our work is the disentanglement between light and texture. This factorization is what allows us to adhere to the photo-consistency assumption, in the sense that different intensities for the same parts of the hand texture can be explained by lighting conditions. Since we are not interested in high-frequency environment mapping, we use spherical harmonics to model the environment light. We employ 12 coefficients to represent the environment map, sufficient to capture most of the information present in natural lighting conditions. The lighting conditions in InterHand2.6M, in particular, are comfortably accommodated in this formulation, as the goal of the developers during capture was to provide uniform lighting.

The use of Redner [30] as our rendering engine enables the correct routing of updates from the loss term to the environment map. This allows us to incorporate the lighting conditions into our optimization process and achieve more accurate results.

3.5. Temporal Prior term

To account for the temporal relationship between the correctives across time t, we use a temporal smoothing term to encourage smoother gesturing. This term is imposed on the articulation correctives δh_{a_t} for any point during optimization and is formulated as follows:

$$L_{prior}\left(\delta h_{a_t}\right) = \left\| h_{a_t} + \delta h_{a_t} - BF\left(h_{a_t} + \delta h_{a_t}\right) \right\|, \quad (4)$$

where $BF\left(\cdot\right)$ applies time-wise Bilateral Filtering (BF) [67] to the input signal. We use BF because we need the articulation correctives to be robust against discontinuities that are due to noise while accommodating greater discontinuities that are likely to be structured.

There are two sources of structured discontinuities. The first is the motion of the user, which could be abrupt. The second and most prevalent source, is the irregular sampling of videos in time in InterHand2.6M. After close inspection, we observed that the videos are quite irregularly sampled in time. Not only that, but each $I_{c,t}$ is not associated with exploitable time information (*i.e*. timestamp), in any way.

The effect that temporal smoothing has on optimization is that it regularizes an ill-posed problem, which otherwise would lead to deviations rather than convergence during optimization. We did not pursue filtering beyond the articulation correctives, as this aspect of hand pose was experimentally determined to suffice.

3.6. Optimization

Our approach involves a 3-stage optimization scheduling process. In the 1st stage, we optimize over the environment map e for 10 iterations. Subsequently, in the 2nd stage, we optimize over the hand texture h_d of dimensions 1024×1024, by enabling the 2 lowest LoDs out of 8, for 10 iterations. Finally, in the 3rd stage, we jointly optimize over the all the aforementioned variables, enabling all hand texture LoDs, as well has the hand shape h_s, the pose correctives δh_{p_t} and the articulation correctives δh_{a_t}, for 80 iterations. This order was experimentally determined to yield good results, essentially starting from coarse appearance and proceeding all the way down to fine details. Throughout all stages, we optimize for color constancy as well. Optimization was performed using the Adam optimizer. The learning rates were experimentally determined to be $l_{h_d} = 0.014$ for the the hand texture, $l_{h_d} = 0.003$ for the

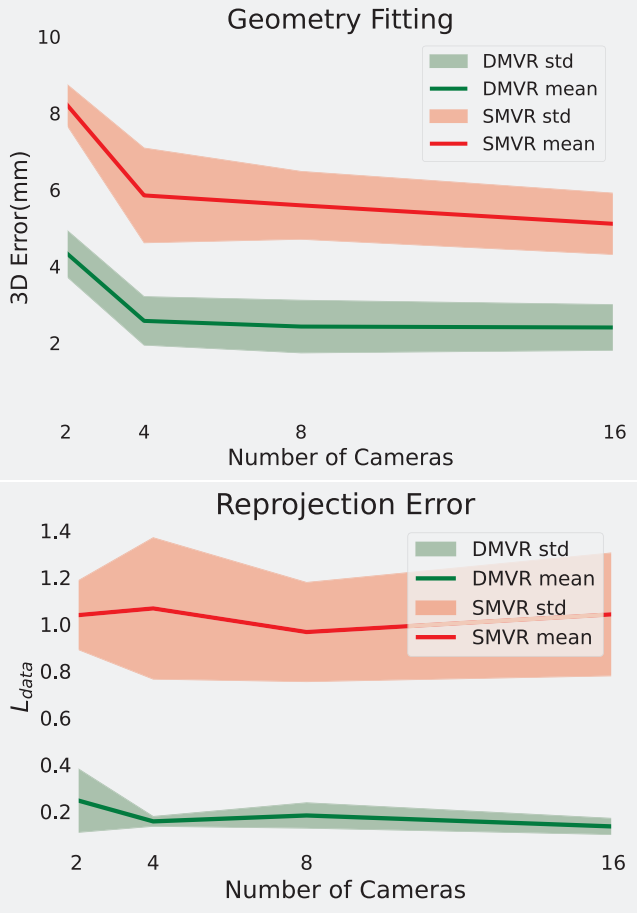


Figure 4: Comparison of the fine-tuning ability, between DMVR and SMVR on synthetic data, against well-defined ground truth. DMVR clearly outperforms SMVR in all aspects. The plots encode mean values and standard deviations of error metrics, with lower values being better.

hand shape, $l_{h_a} = 0.0087$ for the articulation correctives, $l_{p_q} = 0.0015$ for pose rotational correctives, $l_{p_x} = 0.0047$ for pose displacement correctives, $l_{cc} = 0.001$ for color constancy, and $l_e = 0.1$ for the environment map. The objective term weights were set to $\lambda_{data} = \lambda_{prior} = 1$. For the BF, we use $\sigma = 0.6$, $a = 0.5$, and the window size is set to 3.

4. Experiments

We compare our method with the method proposed in [24] (SMVR), which we designate as the baseline. We demonstrate on the InterHand2.6M that the SOTA is both quantitatively and qualitatively improved.

4.1. Quantitative Evaluation

We shortly analyze the structure of InterHand2.6M across axes that are relevant to our experimentation. InterHand2.6M contains several calibrated multi-view videos of hand gestures (we focus on single-hand gestures in this work). The recording cameras are distributed across a half-sphere with the hand being at the center of this volume. The videos regard 28 subjects performing more than 20 gestures, each. The videos vary in the count of samples in time. Close inspection reveals that the videos have varying rates and are not timestamped. All cameras use a common reference frame. In this reference frame, 3D annotations are provided, for each sample, using the MANO representation. These annotations are evidently inaccurate. As demonstrated in [40], it is reasonable to assume that for imagery of the type of InterHand2.6M, such annotations can be automatically provided, with a reconstruction error that is no greater than $10mm$ (a $5mm$ error is claimed in [40]).

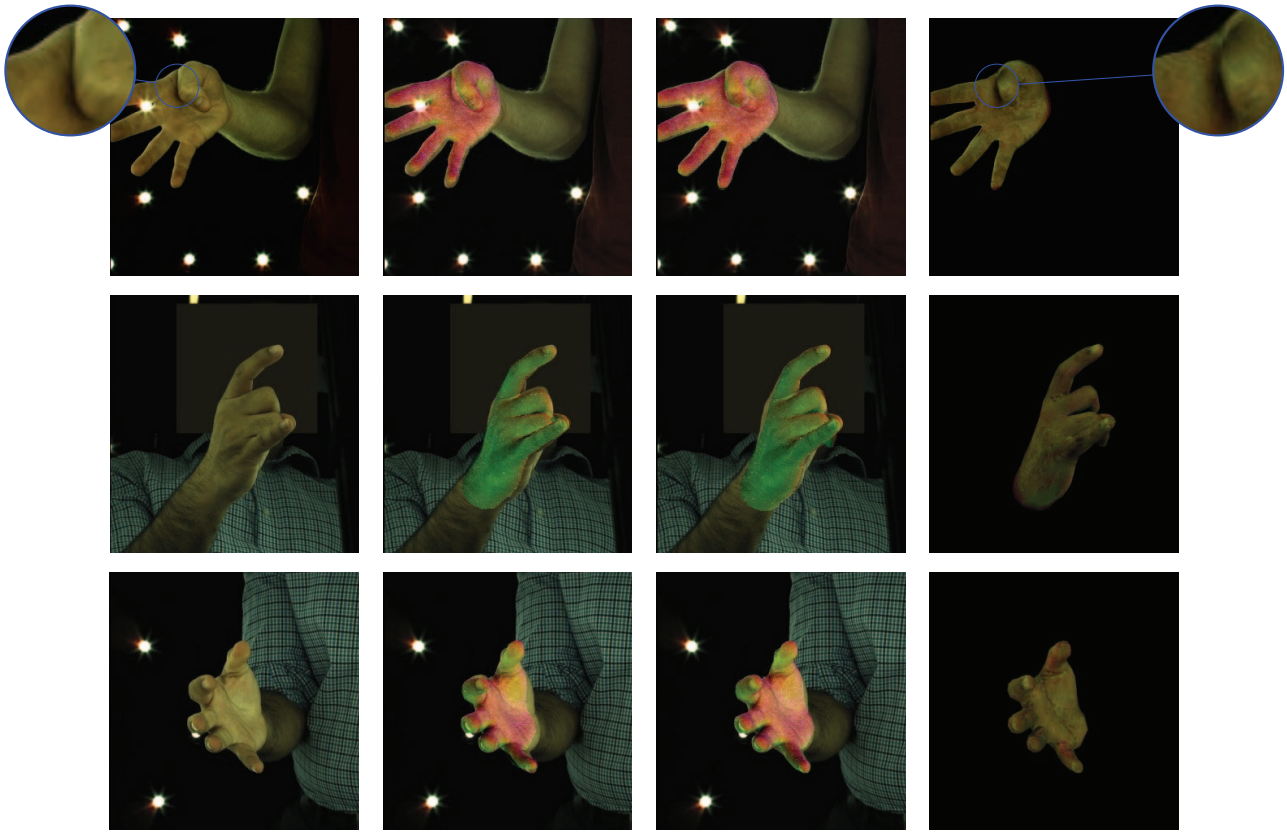
In order to circumvent the lack of reliable ground truth in InterHand2.6M we resort to synthetic experiments, where the ground truth is well-defined. The synthetic data resemble InterHand2.6M by borrowing hand geometry information from the dataset, skinning it with HTML [54] and lighting it artificially. We extend this across time, too.

For any candidate sample (a single hand gesturing video within InterHand2.6M), we synthetically generate a randomized approximation x_0 of the provided ground truth, which is chosen close to the actual ground truth x^* in our synthetic data, and is constrained to have a $10mm$ reconstruction error. The generated imagery, the camera information and x_0 is provided to both DMVR and SMVR. If T is the amount of time samples in the video, then DMVR solves for the entire video jointly, while SMVR solves for each time instant independently. All results are compared with respect to reconstruction error. We vary the camera count $\mathbf{C}$, to estimate the improvement that DMVR has over SMVR as a function of $\mathbf{C}$. The camera selection prioritizes cameras that see larger areas of the hand on average.

The results provided in Fig. 4 indicate that DMVR clearly and always outperforms SMVR in fine-tuning x_0, measured in 3D error. Interestingly, DMVR seems to have a better ability to incorporate the provided visual cues as the number of cameras increases. The improvement in 3D error goes hand-in-hand with an improvement in reprojection error, as shown in the same figure. According to the results, it can be argued that time can be more valuable than adding new camera views. Of course, this is not entirely independent of the actual poses being captured, but is very relevant to the task of designing data acquisition sessions, as in [40].

4.1.1 Hand Texture Reconstruction

DMVR extracts textures that are more faithful to the true textures, compared to SMVR. This is attributed to two factors, namely the integration over time which reveals more of the true texture and the light-texture disentanglement which leads to the right kind of updates on the time-independent



(a) (b) (c) (d)

Figure 5: Exemplar results on InterHand2.6M. The columns represent (a) actual data, (b) bundled annotations, (c) the improvement that DMVR can incur, and, (d) how DMVR reconstructs the relevant observations, through rendering all estimated parts. The better fitting of the normal-mapped silhouettes is revealing of the improvement. It is important to note that the shadows are simulated through ray tracing, are part of the optimization, and are factored out of the extracted texture.

hand texture. Despite the disentanglement, there are still elements of ill-posedness in our formulation, as infinite combinations of color tones and intensities, on the lights and the texture, can yield the same final result. To be robust against this ill-posedness when comparing to ground truth textures, we employ a robust Peak Signal-to-Noise Ratio (PSNR) metric, which we term as $\widehat{\text{PSNR}}$. $\widehat{\text{PSNR}}$ becomes invariant to global color tone and intensity by being estimated up to a 3×4 linear color transform M:

$$\widehat{\text{PSNR}}\left(I_0, I_1\right) = \arg\max_{M} \text{PSNR}\left(I_0, M\left(I_1\right)\right). \quad (5)$$

By computing the improvement r over all synthetic experiments as:

$$r = \widehat{\text{PSNR}}\left(I_{gt}, I_{\text{DMVR}}\right) / \widehat{\text{PSNR}}\left(I_{gt}, I_{\text{SMVR}}\right), \quad (6)$$

with I_{gt} being a ground truth texture, I_{SMVR} being the texture reconstructed by SMVR and I_{DMVR} being the texture reconstructed by DMVR, we estimate the average improvement to be $\bar{r} = 3.84\%$ over a baseline average of $29dB$. For the combined improvement, that incorporates the geometry estimation, as well as the estimation of the environment lighting, the hand texture, and the comparison of the backprojection errors corresponding to SMVR and DMVR (see Fig. 4).

4.2. Qualitative Evaluation

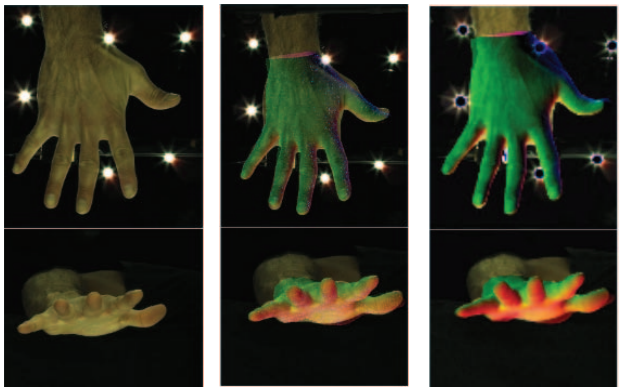
We apply DMVR on real samples, drawn directly from InterHand2.6M, to provide supporting evidence that what is being reported to be quantitatively better than the SOTA transfers to real data, too. For more results please refer to the supplementary.

Analyzed in isolation, DMVR systematically incurs significant geometry refinements on InterHand2.6M, as shown in Fig. 5. The improvement is such that allows a high-frequency reconstruction of the input (see Fig. 5d), which could not be achieved for geometry estimations that do not align well across views, as is the case with the bundled annotations (see Fig. 5b).

Analyzed in comparison to SMVR [24], Fig. 6 acts as

Features	Error	Deterioration
L_{edges} , L_{temp}	**1.962 ± 0.4948mm**	-
L_{temp}	$2.01 \pm 0.4801mm$	2.44%
L_{edges}	$2.032 \pm 0.5394mm$	3.56%
None	$2.41 \pm 0.7845mm$	22.83%

Table 1: Ablation results. Evidently, employing all features is best.



(a) Observation (b) DMVR (c) SMVR

Figure 6: Comparison between SMVR and DMVR on real data. DMVR is evidently advantageous to SMVR in the geometry refinement task.

evidence of the performance improvement that DMVR incurs. The improvement in Fig. 6 is easily discernible at the boundaries, where differences in pixels can amount to differences in millimeters. This difference is associated with a decrease in the number of cameras (32 for SMVR, 13 for DMVR) and the increase in the number of samples in time considered ($\mathbf{T} = 1$ for SMVR, $\mathbf{T} = 10$ for DMVR). The consideration of temporal information also led to the establishment of a full hand texture (see Fig. 7) which is not attainable for SMVR, even in the case of the open hand. We expect even better results for videos that are more revealing, through more complex hand motions.

4.3. Ablation Study

We conducted an ablation study using 16 cameras, to assess the impact of various features on the performance of our method (Table 1). Using all features yielded the best results. Incorporating both L_{edges} and L_{temp} resulted in an error of 1.962mm. Removing L_{temp} caused a slight decline in performance and excluding L_{edges} led to a 3.56% deterioration. Not using any features significantly increased the error by 22.83%. These findings emphasize the critical role of both L_{edges} and L_{temp} in achieving accurate results.

5. Discussion

In this work, we proposed a significant improvement over the SOTA in reconstructing the geometry and appearance of hands from multi-view video sequences. Our method combines temporal consistency with path tracing, leading to accurate and detailed results. We demonstrated the superiority of our method over SOTA on synthetic data, as well as on real-world examples from the InterHand2.6M. Our approach also outperformed existing methods in terms of texture reconstruction.

Figure 7: An example of a texture that DMVR extracted from real data, using a sequence of $\mathbf{T} = 7$ multi-views. A full texture has been recovered despite self-occlusions and occlusions from the acquisition equipment itself.

We are thrilled about the possibilities of enhancing well-established and widely recognized datasets by improving their accuracy and extracting additional relevant information, such as environment lighting and hand textures. This opens the door for additional future improvements. For example, we are interested in endowing our modeling with physical substance, in order to more effectively handle more difficult poses that require the incorporation of stronger priors to properly infer visually ambiguous situations (*e.g.* fist). Incorporating physical substance will help us extend our method to the case of two interacting hands. We are also interested in applying our improvements to more similar datasets (see Section 2.2) and to improve the computational cost, the rendering (*e.g.* the latest installment of the Mitsuba renderer [21] shows promise) and the optimization dynamics. Finally, an interesting direction of further improvement is to extend our framework to jointly consider multiple videos at once, so that the observation of time-independent variables and its robustness, are further improved.

Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of Quadro P6000 and Titan V GPUs used for the execution of this research. This research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for H.F.R.I Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, number 91.

References

[1] Anima Anandkumar, Gil Press, and Andrew Ng. Data collection and quality challenges in deep learning: A data-centric approach. *arXiv preprint arXiv:2112.06409*, 2021. 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2, 3

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 3

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, Nadia Magnenat-Thalmann, and Junsong Yuan. 3d hand pose estimation using synthetic data and weakly labeled rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2

[5] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 2

[6] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20544–20554, 2022. 2

[7] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. *arXiv preprint arXiv:2211.12782*, 2022. 3

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. *arXiv preprint arXiv:2103.11703*, 2021. 3

[9] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. 2, 3

[10] Martin de La Gorce, David J Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1793–1805, 2011. 2

[11] Martin de La Gorce, Nikos Paragios, and David J Fleet. Model-based hand tracking with texture, shading and self-occlusions. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 2

[12] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *arXiv preprint arXiv:2204.13662*, 2022. 3

[13] Yafei Gao, Yida Wang, Pietro Falco, Nassir Navab, and Federico Tombari. Variational object-aware 3-d hand pose from a single rgb image. *IEEE Robotics and Automation Letters*, 4(4):4239–4246, 2019. 2

[14] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation from single depth images using multi-view cnns. *IEEE Transactions on Image Processing*, 27(9):4422–4436, 2018. 2

[15] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2

[16] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Handsformer: Keypoint transformer for monocular 3d pose estimation of hands and object in interaction. *arXiv preprint arXiv:2104.14639*, 2021. 2

[17] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (TOG)*, 2020. 2

[18] Shangchen Han, Beibei Liu, Robert Wang, Yuting Ye, Christopher D Twigg, and Kenrick Kin. Online optical marker-based hand tracking with deep labels. *ACM Transactions on Graphics (TOG)*, 2018. 2

[19] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 3

[20] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 2

[21] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, Merlin Nimier-David, Delio Vicini, Tizian Zeltner, Baptiste Nicolet, Miguel Crespo, Vincent Leroy, and Ziyi Zhang. Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org. 8

[22] Michael I Jordan, Jon M Kleinberg, and Sendhil Mullainathan. Economics and ai: An unlikely alliance. In *Proceedings of the National Academy of Sciences*, volume 116, page 13702–13705, 2019. 2

[23] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. *arXiv preprint arXiv:2212.09530*, 2022. 3

[24] Giorgos Karvounas, Nikolaos Kyriazis, Iason Oikonomidis, Aggeliki Tsoli, and Antonis A Argyros. Multi-view image-based hand geometry refinement using differentiable monte carlo ray tracing. In *British Machine Vision Conference (BMVC 2021)*, Virtual, UK, November 2021. BMVA. 1, 2, 4, 5, 6, 7

[25] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2

[26] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Real time hand pose estimation using depth sensors. In *Consumer depth cameras for computer vision*, pages 119–137. Springer, 2013. 2

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[28] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2, 3

[29] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326*, 2019. 3

[30] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics (TOG)*, 37(6):1–11, 2018. 4, 5

[31] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. Piano: A parametric hand bone model from magnetic resonance imaging. *arXiv preprint arXiv:2106.10893*, 2021. 3

[32] Yuncheng Li, Zehao Xue, Yingying Wang, Liuhao Ge, Zhou Ren, and Jonathan Rodriguez. End-to-end 3d hand pose estimation from stereo cameras. In *BMVC*, volume 1, page 2, 2019. 2

[33] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 3

[34] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. Handtailor: Towards high-precision monocular 3d hand recovery. *arXiv preprint arXiv:2102.09244*, 2021. 3

[35] Alexandros Makris and A Argyros. Model-based 3d hand tracking with on-line hand shape adaptation. In *Proc. BMVC*, pages 77–1, 2015. 2

[36] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 380–397. Springer, 2022. 2

[37] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 2

[38] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Neuralannot: Neural annotator for 3d human mesh training sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2307, 2022. 3

[39] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*, pages 440–455. Springer, 2020. 3

[40] 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*, pages 548–564. Springer, 2020. 1, 3, 6

[41] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 2

[42] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1154–1163, 2017. 3

[43] Andrew Ng. Data-centric ai. https://datacentricai.org/, 2021. 2

[44] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Training a feedback loop for hand pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3316–3324, 2015. 2

[45] Takehiko Ohkawa, Yu-Jhe Li, Qichen Fu, Ryosuke Furuta, Kris M Kitani, and Yoichi Sato. Domain adaptive hand keypoint and pixel localization in the wild. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 68–87. Springer, 2022. 2

[46] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Markerless and efficient 26-dof hand pose recovery. In *Asian Conference on Computer Vision*, pages 744–757. Springer, 2010. 2

[47] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, volume 1, page 3, 2011. 2

[48] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 2

[49] Paschalis Panteleris and Antonis Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 575–584, 2017. 2

[50] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018. 2

[51] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1496–1505, 2022. 2

[52] Gil Press. Andrew ng launches a campaign for data-centric ai. *Forbes*, 2021. 2

[53] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1106–1113, 2014. 2

[54] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision*, pages 54–71. Springer, 2020. 3, 6

[55] James M Rehg and Takeo Kanade. Digiteyes: Vision-based hand tracking for human-computer interaction. In *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 16–22. IEEE, 1994. 2

[56] Rilwan Remilekun Basaru, Greg Slabaugh, Eduardo Alonso, and Chris Child. Hand pose estimation using deep stereovision and markov-chain monte carlo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 595–603, 2017. 2

[57] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20555–20565, 2022. 2

[58] Javier Romero, Hedvig Kjellström, and Danica Kragic. Monocular real-time 3d articulated hand pose estimation. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 87–92. IEEE, 2009. 2

[59] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 2017. 2, 3, 4

[60] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)*, 2020. 2

[61] Adrian Spurr, Aneesh Dahiya, Xucong Zhang, Xi Wang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. *arXiv preprint arXiv:2106.05953*, 2021. 2

[62] Adrian Spurr, Pavlo Molchanov, Umar Iqbal, Jan Kautz, and Otmar Hilliges. Adversarial motion modelling helps semi-supervised hand pose estimation. *arXiv preprint arXiv:2106.05954*, 2021. 2

[63] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 2

[64] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer Graphics Forum*, volume 34, pages 101–114. Wiley Online Library, 2015. 2

[65] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014. 2

[66] Peter Thompson and Aphrodite Galata. Hand tracking from monocular rgb with dense semantic labels. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 394–401. IEEE, 2020. 2

[67] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998. 5

[68] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[69] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 2

[70] Jan Wöhlke, Shile Li, and Dongheui Lee. Model-based hand pose estimation for generalized hand shape with appearance normalization. *arXiv preprint arXiv:1807.00898*, 2018. 2

[71] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 122–139. Springer, 2020. 3

[72] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3895–3905, 2022. 2

[73] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 3

[74] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. *arXiv preprint arXiv:2106.04324*, 2021. 2

[75] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. Technical report, arXiv:1705.01389, 2017. https://arxiv.org/abs/1705.01389. 2

[76] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 3