Learn 2D, Solve 3D: UA-Fit, an Uncertainty-Weighted Analytical Solver for Multi-view Hand Mesh Recovery

Marsil Zakour (✉), Constantin Patsch, Martin Piccolrovazzi, Yuankai Wu, and Eckehard Steinbach

Technical University of Munich,
School of Computation, Information, and Technology,
Department of Computer Engineering, Chair of Media Technology,
Munich Institute of Robotics and Machine Intelligence (MIRMI),
Munich, Germany
{marsil.zakour, constantin.patsch, martin.piccolrovazzi, yuankai.wu, eckehard.steinbach}@tum.de

Abstract. In multi-view hand capture, the most accurate estimators regress free mesh vertices. Methods that regress MANO parameters are usually less accurate, since rotations are hard to regress and errors amplify along the kinematic chain, while fitting is accurate but slow. We show that when the cameras and MANO kinematics are known, the lift to 3D needs no learned parameters, as only the 2D evidence is uncertain. Our method UA-Fit (Uncertainty-weighted Analytical Fit) learns only per-view 2D evidence: a dense vector field where each patch of pixels votes for each joint, with an anisotropic precision per vote. A parameter-free solver recovers MANO pose and translation: it triangulates the votes into precision-weighted 3D joints, then fits MANO with a batched Levenberg–Marquardt solver whose analytic kinematic Jacobian is ≈14× cheaper than automatic differentiation. The hand shape is predicted by a compact feed-forward head. The solver has no learned weights, yet we train the 2D fields end to end through its converged solution at constant memory. The output is MANO-consistent by construction, and the learned reconstruction stage is only 2.0M parameters, 20× smaller than the learned fusion head of POEM-v2's parametric variant (1.6× smaller in total). On the full evaluation splits of five benchmarks under one protocol, UA-Fit averages 6% lower per-vertex and 11% lower per-joint error than this baseline: markedly more accurate on the shape-diverse InterHand2.6M and HO3D, on par on ARCTIC, and less accurate on DexYCB and OakInk, where the 2D evidence, not the solver, is the limit. Zero-shot results on GigaHands and an internal dataset support generalization. Code and project page: https://marcilzakour.github.io/ua-fit/.

Keywords: multi-view hand pose · MANO · analytical solver · efficiency · uncertainty

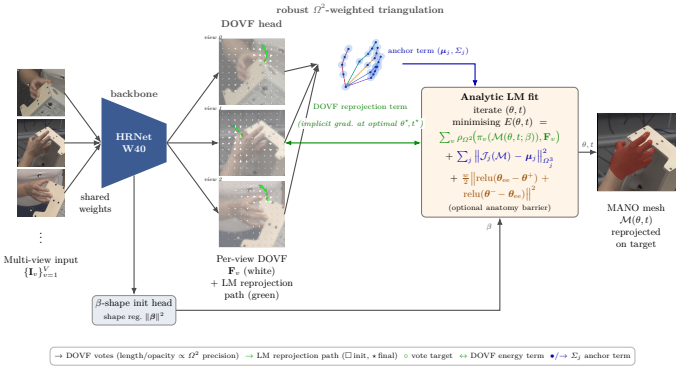


Fig. 1: UA-Fit pipeline. A shared HRNet backbone maps the V input views to a per-view *dense offset vector field* (DOVF) f_v with anisotropic precision $\boldsymbol{\Omega}$, and to a feed-forward $\boldsymbol{\beta}$-shape head. Precision-weighted robust triangulation lifts the votes to 3D joints with closed-form covariance $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$. A parameter-free analytic Levenberg–Marquardt (LM) solver then fits MANO $(\boldsymbol{\theta}, \mathbf{t})$ by minimising an energy with a DOVF reprojection term (green), an $\boldsymbol{\Omega}$-weighted 3D-anchor term (blue), and a pose prior. The optional anatomy barrier (orange) is shown. Gradients reach the 2D evidence through the LM fixed point by implicit differentiation at the optimum $(\boldsymbol{\theta}^\star, \mathbf{t}^\star)$. The triangulation and LM solver contain *zero* trainable parameters.

1 Introduction

Calibrated multi-camera rigs are standard for hand motion capture, resolving the depth and occlusion ambiguities that limit single-view estimators [9, 41, 44] in virtual- and augmented-reality (VR/AR) labs, robot teleoperation, and dataset collection [3,6,8,34]. In this calibrated setting the most accurate models are *non-parametric*, regressing free vertices directly [27–29, 63] rather than the parameters of a hand model. Predicting parameters is harder. Direct MANO [48] pose regression is harmed by rotation-representation discontinuities [69], kinematic-chain error amplification [25], and broken image alignment [32, 66]. The clearest evidence is head-to-head: POEM-v2 [63], the accuracy reference for this setting, releases a MANO-parametric variant and a free-vertex model *of the same size*, and the parametric one is less accurate on every benchmark (on HO3D [12], 9.16 vs 7.63 mm MPVPE).

Yet many applications (animating hands, retargeting to robots, reasoning about contact) need the low-dimensional parametric MANO representation, and recovering it afterward by fitting MANO to predicted vertices is itself a lossy optimization [1, 19]. Accuracy also carries a compute cost: the dominant recipe pairs a shared backbone with a *large learned fusion head* that aggregates views in 3D [29, 61, 63], in POEM-v2 a point-embedded transformer with 40.3M parameters (parametric variant) or 92.8M (largest vertex model). The accuracy leaders are thus both non-parametric and heavy, which many deployments cannot accept.

This paper starts from a simple observation: *in the calibrated setting, the 2D-to-3D lift need not be learned.* Camera geometry is known exactly, and so is the hand's kinematic structure (MANO); what is uncertain, and hence worth

learning, is the *2D image evidence and its reliability*. We therefore ask how small the learned part of a multi-view hand estimator can be if geometry and the MANO prior do all the 3D work.

Our answer, UA-Fit (Fig. 1), learns only a per-view 2D front-end: a *dense offset vector field* (DOVF), where every pixel votes for every joint, and a per-pixel 2×2 Cholesky-parameterized *anisotropic precision field* $\boldsymbol{\Omega}$, trained with a Gaussian negative log-likelihood so the network reports the directional uncertainty of its own votes. Beyond this front-end and a compact feed-forward shape head, everything downstream is classical estimation with *zero learned parameters*: uncertainty-weighted robust triangulation yields metric 3D joints with a closed-form covariance, and a Kabsch-initialized [16], block-weighted Levenberg–Marquardt solver fits MANO to the multi-view 2D fields (Sec. 3). The components are deliberately classical; the contribution is their interface: a dense 2D likelihood whose full 2×2 anisotropic precision is propagated in closed form through triangulation into the articulated fit, and is itself trained through the converged solver.

This runs counter to the prevailing scaling trend: where recent work grows backbones and fusion heads for incremental accuracy [29,63], our learned reconstruction stage is 2.0M parameters, 20× smaller than the parametric fusion head it replaces (1.6× smaller in total, since the two methods share the backbone). Across five calibrated benchmarks it matches POEM-v2's trained parametric variant on average, most clearly on the shape-diverse InterHand2.6M set where free parameter regression degrades, and transfers best in zero-shot evaluation (Sec. 4.2).

We make three contributions:

1. **A parametric multi-view hand estimator with no learned 3D fusion.** We replace the large learned fusion head with a small per-view 2D evidence network and a parameter-free solver, yielding a MANO-consistent hand by construction that matches POEM-v2's trained parametric variant on average across five benchmarks (Table 1).
2. **An uncertainty-weighted analytical solver.** We cast MANO recovery as optimization directly against the learned 2D fields, with no learned parameters. Its analytic articulated Jacobian closely matches automatic differentiation at ≈14× lower cost, an implicit backward trains the 2D fields through the solver at $O(1)$ memory, and the explicit energy composes new terms at inference. A controlled solver study isolates each choice (Sec. 4.3).
3. **Learned reliability, not learned capacity, is what matters.** The learned anisotropic precision $\boldsymbol{\Omega}$ is what makes the parameter-free solver competitive: removing it ($\boldsymbol{\Omega}=I$, same weights) degrades every benchmark (Sec. 4.3). Capacity in the 3D fusion stage is not what is missing: our zero-parameter solver matches POEM-v2's 40.3M learned fusion head in distribution and generalizes better on two datasets neither model was trained on (Table 1).

More capacity still helps: the non-parametric POEM-v2-large stays up to ∼2 mm ahead at high view counts (Sec. 5), and such capacity can enter our solver as energy terms, whether as sharper 2D evidence from a stronger backbone or as

2D and 3D predictions from larger, even non-parametric, models, which we leave to future work. Our claim is narrower: with calibrated rigs, parametric output, and limited compute, classical estimation with learned uncertainty makes a large learned 3D stage unnecessary.

2 Related Work

Monocular hand mesh recovery. Single-image hand estimation has grown from RGB keypoint regression [70] into full mesh recovery, splitting into two families. *Model-based* (parametric) methods regress MANO [48] pose and shape and decode them through the differentiable MANO layer [2,39,67], guaranteeing an anatomically plausible hand but inheriting the pitfalls of parameter regression (below). *Model-free* (non-parametric) methods instead predict vertices or 2.5D directly through graph-convolution, volumetric, registration, spiral, and transformer decoders [4, 5, 9, 22, 27, 28, 32, 33], all trading guaranteed validity for tighter image alignment. Recent large-scale vision-transformer (ViT) systems return to parametric output (HaMeR [41], WiLoR [44]), and two-hand methods add explicit interaction reasoning [26, 31]. All resolve depth and scale from a single view; we instead exploit calibrated multi-view geometry, so the network never learns the 2D-to-3D lift.

Multi-view hand mesh recovery. POEM [61] and POEM-v2 [63] embed basis points in the intersection of camera frusta and let mesh queries cross-attend to multi-view features, defining the current accuracy reference for calibrated generalizable hand mesh recovery. The large variant couples an HRNet [50] backbone with a 92.8M point-embedded transformer. MLPHand [59] targets real-time multi-view reconstruction with per-view MLPs, and HGGT [29] scales a VGGT/DINOv2 [38, 55] foundation transformer, part of a broader move toward feed-forward multi-view 3D reconstruction [56], to *uncalibrated* multi-view hands, evaluated only under root-relative or Procrustes alignment. The common thread is a *large learned fusion head* performing 3D aggregation inside the network; we instead remove that stage and compare under identical calibrated protocols against POEM-v2's released models.

Geometric multi-view fusion and triangulation. A parallel line for human *bodies* keeps 3D reasoning geometric: confidence-weighted direct linear transform (DLT) and learnable triangulation backpropagate through a weighted singular-value decomposition [15], AdaFuse [68] learns per-view fusion weights, cross-view and epipolar attention share evidence before lifting [14, 46], and aggregators fuse in a canonical 3D grid or camera-disentangled space [47, 54]. These weight *point* observations with *scalar* confidences inside an otherwise fixed triangulation, echoing classical robust geometric estimation such as bundle adjustment, RANSAC, and PnP [7, 23, 53]. We generalize the evidence to dense offset fields with full 2×2 anisotropic precisions, propagate them to a closed-form 3D covariance, and continue past triangulation into a robust articulated model fit, where validity and parametric output arise.

Parametric models, regression and optimization-based fitting. Parametric shape models (MANO [48] for hands, SMPL/SMPL-X [30,40] for bodies) underpin most mesh recovery, either by direct regression [17,20] or by optimization. Direct regression of MANO parameters is harmed by rotation-representation discontinuities [69] and by kinematic error amplification along the chain. HybrIK [25] addresses this for bodies with a hybrid analytical IK, and pixel-aligned methods [32, 66] document why parameter regression breaks image alignment. The optimization tradition instead *fits* the model to image evidence, from SMPLify [1] and its reprojection-energy successors to in-the-loop fit-and-regress [19], amortized learned-gradient and neural-descent solvers [49, 64], and probabilistic fitting over solution sets [21]. We avoid regression entirely: parameters are obtained by *optimization* against learned, uncertainty-weighted image evidence, with the MANO manifold as the regularizer. General differentiable-optimization libraries such as Theseus [43] offer autograd LM solvers, which we compare against in the solver study (Sec. 4.3).

Uncertainty and dense evidence representations. Heatmap regression is the dominant 2D-localization paradigm [35,50,57], later refined by integral/soft-argmax decoding [36,51], sub-pixel distribution modeling [10,65], and ViT backbones [58]. Predicting per-pixel *votes* rather than a single peak underlies dense-correspondence and six-degree-of-freedom (6-DoF) object-pose voting networks [11, 42], which inspire our dense offset field. Orthogonally, modeling *uncertainty* (aleatoric and epistemic [18], directional-statistics heads [45], and residual log-likelihood regression [24]) calibrates how much each prediction should be trusted. We unify these threads: every pixel votes for every joint (dense evidence) *and* reports an anisotropic 2×2 precision (calibrated uncertainty); unlike prior work that stops at 2D or at scalar-weighted triangulation, we propagate that precision analytically through triangulation and the MANO fit.

Datasets. We evaluate on five standard calibrated multi-view benchmarks: DexYCB [3], HO3D [12], ARCTIC [6], InterHand2.6M [34], and OakInk [60], and report zero-shot transfer to GigaHands [8]. This ecosystem [13,37,52,71] increasingly demands *parametric*, animation-ready output, the regime our MANO-consistent fit targets.

3 Method

UA-Fit recovers MANO parameters from V calibrated views through one learned stage and a 3D stage whose solver is parameter-free (Fig. 1). The learned stage is a shared backbone that predicts, per view, *dense 2D evidence*: where every joint projects and how reliable that vote is (Sec. 3.2). The 3D stage's geometry carries no learned weights: it lifts this evidence to metric 3D joints with a closed-form covariance by precision-weighted robust triangulation (Sec. 3.3), then initializes a canonical MANO hand at those joints and fits its pose to the same evidence with an analytic Levenberg–Marquardt solver anchored by that covariance. The MANO shape is predicted feed-forward by a small learned head (Sec. 3.4). The

output is a point on the MANO manifold, MANO-consistent by construction, whose accuracy is bounded by the 2D evidence.

3.1 Preliminaries and notation

Problem. We are given V calibrated views of a single right hand (crops $I_v \in \mathbb{R}^{3\times256\times256}$ with known intrinsics K_v and world-to-camera extrinsics T_v, each roughly centered on the hand) and recover a parametric MANO [48] hand: pose $\boldsymbol{\theta} \in \mathbb{R}^{48}$, shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and global translation $\mathbf{t} \in \mathbb{R}^3$. Intrinsics and extrinsics are treated as exact. We assume one hand per crop, with left hands mirrored to the canonical right-hand MANO convention in the data loader.

The MANO hand model. MANO maps $(\boldsymbol{\theta}, \boldsymbol{\beta})$ to a mesh by linear blend skinning of a template deformed by shape and pose blendshapes,

$$M(\boldsymbol{\theta}, \boldsymbol{\beta}) = W\big(\bar{T} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}),\ \mathcal{J}(\boldsymbol{\beta}),\ \boldsymbol{\theta},\ \mathcal{W}\big), \tag{1}$$

with skinning weights $\mathcal{W}$ over a kinematic tree of 16 joints. The pose $\boldsymbol{\theta}$ splits into a global rotation and 15×3 axis-angle articulations. The shape blendshape B_S and joint regressor $\mathcal{J}(\boldsymbol{\beta})$ fix bone lengths, and the pose-corrective blendshape $B_P(\boldsymbol{\theta})$ absorbs skinning artifacts. Forward kinematics over the tree gives each joint j a global rotation R_j^g and origin $\mathbf{o}_j$, from which the 778 vertices and J=21 OpenPose-convention joints $\mathbf{x}_j$ follow. We fit $\boldsymbol{\theta}, \mathbf{t}$ (and optionally $\boldsymbol{\beta}$). The two structures the analytical solver exploits are the kinematic tree (for an analytic articulated Jacobian, Sec. 3.4) and $\mathcal{J}(\boldsymbol{\beta})$, for bone-length recovery.

Multi-view geometry. A world point $\mathbf{x}$ projects into view v as $\mathbf{u} = \pi_v(\mathbf{x}) = \Pi(K_v T_v \mathbf{x})$, with Π the pinhole division. Seen in ≥ 2 views, a point is recovered by *triangulation*: the direct linear transform (DLT) gives a closed-form least-squares intersection of the back-projected rays, which we refine (Sec. 3.3). At two views the solution is weakly constrained along the inter-camera baseline, a property that reappears as an anisotropic 3D covariance (Sec. 5).

Estimation in information form. Throughout, per-observation reliability is a *precision* matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ and every data term is a squared Mahalanobis residual $\mathbf{r}^\top \boldsymbol{\Omega}\, \mathbf{r}$. Minimizing a sum of such terms is nonlinear weighted least squares, solved by Gauss–Newton (GN) with Levenberg–Marquardt (LM) trust-region damping and iteratively reweighted least squares (IRLS) for outliers. These are the only tools in the 3D stage, and their MANO-specific forms are given in the supplement.

3.2 Learned 2D evidence: dense offset fields with anisotropic precision

A shared HRNet-W40 backbone [50] processes each view independently at 256^2 input (resized from the stored crops, Sec. 3.5) on a 64^2 feature/heatmap grid. A light feature neck, the per-view heads, and a small scene-level shape head (Sec. 3.4) are the *only* learned reconstruction components (2.0M parameters

in total). All downstream 3D estimation is parameter-free. The per-view heads produce:

Dense offset vector field (DOVF). $f_v \in \mathbb{R}^{J\times 2\times 64\times 64}$, where $f_v[j](\mathbf{p})$ is the 2D offset from pixel $\mathbf{p}$ to joint j. Unlike a heatmap, whose information collapses to its argmax, the field is a *distributed, redundant* encoding: every pixel in the support votes, and the field remains informative at locations *off* the true joint, the property the solver exploits (Sec. 3.4).

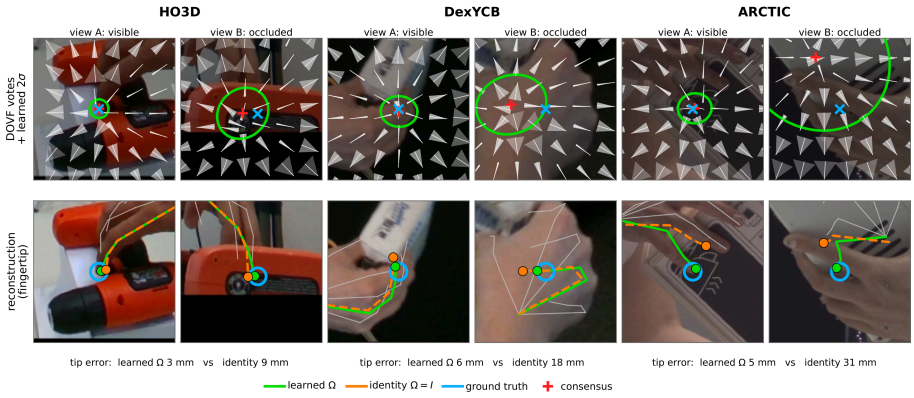


Fig. 2: Learned uncertainty on two-view examples (HO3D, DexYCB, ARCTIC). A fingertip is visible in view A and occluded in view B. **Top:** DOVF votes with the learned 2σ ellipse ($\boldsymbol{\Sigma}{=}\boldsymbol{\Omega}^{-1}$, green), tight where visible and wide under occlusion. **Bottom:** weighting the two-view triangulation by the learned $\boldsymbol{\Omega}$ downweights the occluded view and lands the tip on the ground truth (cyan ring), whereas identity weighting drifts several-fold further (errors inset). The solver weights every 2D residual by this $\boldsymbol{\Omega}$ (Sec. 4.3).

Anisotropic precision field. A three-channel Cholesky field $L_v \in \mathbb{R}^{J\times 3\times 64\times 64}$ parameterizes a symmetric positive-definite 2×2 precision $\boldsymbol{\Omega}_v[j](\mathbf{p}) = LL^\top$ per pixel and joint. It is trained with the Gaussian negative log-likelihood (NLL) of the field residuals,

$$\mathcal{L}_{\text{nll}} = \tfrac{1}{2}\,\mathbf{r}^\top \boldsymbol{\Omega}\,\mathbf{r} - \tfrac{1}{2}\log\det\boldsymbol{\Omega}, \tag{2}$$

so the network learns to report *directional* confidence: tight ellipses on well-localized joints, elongated along the ambiguous direction under occlusion, edge-parallel ambiguity, or blur (Fig. 2).

2D consensus. At inference, each pixel $\mathbf{p}$ casts a vote $\mathbf{p} + f_v[j](\mathbf{p})$ for joint j. Voting over the field yields a per-joint consensus location $\mathbf{u}_{vj}$ (the $\boldsymbol{\Omega}$-weighted mean of the votes) with an aggregated anisotropic precision $\boldsymbol{\Omega}_{vj}$, so that agreement across many pixels sharpens the estimate. The precision field's NLL (Eq. (2)) is calibrated *through the solver* under 3D supervision (Stage B, Sec. 3.5), as its likelihood needs 3D grounding rather than the field's own vote scatter.

3.3 $\boldsymbol{\Omega}$-weighted robust triangulation with closed-form covariance

Given the per-view consensus points and precisions $(\mathbf{u}_{vj}, \boldsymbol{\Omega}_{vj})$ from Sec. 3.2, the 3D stage's first step estimates metric 3D joints $\boldsymbol{\mu}_j$ by minimizing their precision-

weighted reprojection error,

$$\boldsymbol{\mu}_j = \arg\min_{\mathbf{x}} \sum_v \mathbf{r}_{vj}^\top \boldsymbol{\Omega}_{vj} \mathbf{r}_{vj}, \qquad \mathbf{r}_{vj} = \mathbf{u}_{vj} - \pi_v(\mathbf{x}), \tag{3}$$

solved with two Gauss–Newton steps from the DLT solution. Because Eq. (3) is a weighted least-squares problem, the inverse GN Hessian directly provides a per-joint 3D covariance, $\boldsymbol{\Sigma}_j = \left(\sum_v J_{vj}^\top \boldsymbol{\Omega}_{vj} J_{vj}\right)^{-1}$, with $J_{vj} = \partial \pi_v / \partial \mathbf{x}$ the projection Jacobian (propagation derived in the supplement), which we retain: it tells the downstream fit *how much to trust the triangulation, per joint and per direction* (*e.g.*, elongated along the dominant baseline at low view counts).

Robustness. Triangulation reweights views by a redescending Cauchy IRLS on the Mahalanobis residual, which drives to zero the weight of a view that localizes the wrong hand whenever at least two views agree (supplement). Under our protocol, where each crop is centred on the target hand, it is accuracy-neutral, and we keep it as insurance for the detector-in-the-loop setting we scope out (Sec. 5).

3.4 Analytical MANO fitting to dense field energies

With the triangulated joints and covariances $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ as initialization and anchor, the second stage fits MANO directly to the learned 2D fields, the core of UA-Fit. Let $\mathbf{x}_j(\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{t})$ denote the model joints. We minimize

$$E(\boldsymbol{\theta}, \mathbf{t}) = \underbrace{\sum_{v,j} \mathbf{r}_{vj}^\top s_{vj} \boldsymbol{\Omega}_{vj} \mathbf{r}_{vj}}_{\text{dense field energy}} + \underbrace{\lambda_{3\text{D}} \sum_j \tilde{\mathbf{r}}_j^\top \boldsymbol{\Sigma}_j^{-1} \tilde{\mathbf{r}}_j}_{\text{3D anchor}} + \underbrace{\mathcal{P}(\boldsymbol{\theta})}_{\text{pose prior}}, \tag{4}$$

with three properties that distinguish it from classical point-based refitting:

(i) The residual is read from the field, not from a fixed point. $\mathbf{r}_{vj} = f_v[j]\big(\pi_v(\mathbf{x}_j)\big)$ samples the offset field *at the current model projection*: the energy landscape is as wide as the field's support, rather than a quadratic bowl around a single argmax. Every LM iteration re-reads the field (and the precision below) where the model currently is. The Jacobian keeps this dependence, chaining the field's bilinear spatial gradient with the projection and articulated Jacobians, $\partial \mathbf{r} / \partial(\boldsymbol{\theta}, \mathbf{t}) = \nabla_{\mathbf{u}} f \cdot \partial \pi_v / \partial \mathbf{x} \cdot \partial \mathbf{x}_j / \partial(\boldsymbol{\theta}, \mathbf{t})$ (full chain in the supplement).

(ii) Every residual is block-weighted by learned precision. $\boldsymbol{\Omega}_{vj}$ is sampled from the Cholesky field at the same location, so directions the network marked unreliable are automatically discounted. The scalar s_{vj} combines a Huber weight (radius 1.5 heatmap-px) and the view validity, a gentler down-weighting than the triangulation's redescending Cauchy kernel (Sec. 3.3). The 3D anchor $\tilde{\mathbf{r}}_j = \mathbf{x}_j - \boldsymbol{\mu}_j$ is weighted by the triangulation precision $\boldsymbol{\Sigma}_j^{-1}$ with a small $\lambda_{3\text{D}}$, unit-matched to the pixel-scale field energy so that it binds (Sec. 4.3). The prior $\mathcal{P}$ is a PCA-Mahalanobis pose prior anchored to the initialization.

(iii) The solver is a batched, adaptively damped LM with an analytic kinematic Jacobian. The fit starts from a Kabsch alignment [16] of the

canonical MANO skeleton to $\boldsymbol{\mu}$, which resolves the global-rotation ambiguity. From a cold start (identity global rotation) the fit converges to a wrong basin and MPVPE roughly triples (7–9 $\rightarrow$ 21–24 mm across the benchmarks). Each iteration then builds the normal equations

$$A = \sum_{v,j} J^\top \boldsymbol{\Omega} J + \mu I + P, \qquad g = \sum_{v,j} J^\top \boldsymbol{\Omega}\, \mathbf{r} + p, \tag{5}$$

over the 51 parameters $(\boldsymbol{\theta}, \mathbf{t})$, where μ is the per-sample trust-region damping and P, p the pose-prior Gauss–Newton block. The Jacobian uses the *analytic articulated* form $\partial \mathbf{x}_i / \partial \theta_j = \mathbf{a}_j \times (\mathbf{x}_i - \mathbf{o}_j)$, with $\mathbf{a}_j$ the rotation axis and $\mathbf{o}_j$ the centre of joint j, over the kinematic tree with the SO(3) right-Jacobian correction (full chain in the supplement). We deliberately omit the MANO *pose-corrective blendshape* term, as it contributes little to the Gauss–Newton direction. Sec. 4.3 shows that the resulting analytic Jacobian agrees with the exact autograd Jacobian to cosine 0.998, at $\approx$14$\times$ lower cost. We run a fixed budget of 15 iterations, where each forms and solves one dense 51$\times$51 system batched across the scene, costing a few milliseconds per hand. An implicit fixed-point backward makes this fit differentiable end-to-end at $O(1)$ memory. Stage B uses it to train the released model *through* the solver (Sec. 3.5).

Composable inference-time energies. Because E is an explicit objective, further terms add at test time without retraining. One we use is the anatomical-validity barrier E_{anat}, a one-sided squared penalty keeping the anatomy-aligned twist/spread/bend angles of the finger joints inside the published per-axis limits of the A-MANO decomposition [62]. Inside the anatomical box the term is zero, so it never pulls a valid pose off the data optimum. It enters as one extra detached Gauss–Newton block in the last few LM iterations, at inference only, and gives a small consistent gain (Sec. 4.3, with the formula and limits in the supplement).

Shape: feed-forward $\hat{\boldsymbol{\beta}}$. Shape is predicted feed-forward by a compact head ($\sim$70k parameters) combining the scene-pooled image descriptor with the metric bone lengths of the triangulated $\boldsymbol{\mu}$ (trained in Stage B through the fit), so the pipeline runs a *single* LM solve with no test-time shape optimization. The bone-length cue supplies the subject-scale information the pooled descriptor alone cannot recover on shape-diverse two-hand data (the source of the InterHand gain).

Why parametric-by-optimization? The output is a point on the MANO manifold, so the gross failures of free-vertex decoders (broken, non-hand meshes on out-of-distribution shapes, Sec. 4.2) are unrepresentable. MANO-consistency does not by itself guarantee anatomical joint-limit validity, which the optional inference-time anatomy barrier addresses (Sec. 4.3). The fit-to-triangulation mesh floor is only 2.1–2.5 mm, so accuracy is governed by the learned 2D evidence, exactly where the parameters were spent.

3.5 Training setup

UA-Fit is trained in two stages on multi-view crops of a six-dataset mix [3, 6, 12, 34, 60, 71], the same training data as the POEM-v2 models we compare against

Table 1: Parametric multi-view hand recovery (mm, ↓). Both released checkpoints are evaluated on the full official splits at fixed maximum views under one identical protocol following POEM-v2, with mesh-regressed joints and oracle InterHand crops (MANO-parametric methods only, Sec. 4.1). UA-Fit uses one global configuration without per-dataset tuning. **Bold** marks the better method per metric, and PA denotes Procrustes alignment. *Zero-shot*: no fine-tuning (3 LM iterations). §internal multi-view set, right hands (POEM-v2-param does not accept our left-hand mirroring), full split (≈12k frames). ¶GigaHands, full official test set (2026 clips), 6 views, heavy-tailed (medians 14.0 vs **10.0**).

	POEM-v2-param			UA-Fit (ours)		
dataset (views)	MPVPE	MPJPE	PA-MPVPE	MPVPE	MPJPE	PA-MPVPE
In-distribution						
DexYCB (8v)	**7.0**	**7.0**	**4.3**	7.9	7.5	4.9
HO3D (5v)	9.1	8.8	5.7	**8.1**	**7.5**	**5.1**
ARCTIC (8v)	7.6	7.1	**5.6**	**7.5**	**6.6**	5.7
InterHand (8v)	12.0	11.6	7.9	**8.8**	**7.8**	**6.2**
OakInk (4v)	**8.8**	**8.6**	**4.9**	9.6	8.9	5.7
Zero-shot						
Internal (4v)§	16.9	16.9	8.4	**11.5**	**11.3**	**6.7**
GigaHands (6v)¶	16.8	16.0	8.1	**13.7**	**13.2**	**7.5**

(a matched comparison). **Stage A** pre-trains the backbone and DOVF head on 2D targets alone (field regression and an auxiliary heatmap loss, all views per scene, 10 epochs). **Stage B** then adds the precision field, the shape head, and the full LM fit inside the training loop, initialized from Stage A (6 epochs at 2–8 random views). The fitted joints are supervised, and gradients reach the fields and precisions through the implicit fixed-point backward (derived in the supplement). This stage trains the fields, the precision, and the shape head through the solver, and we report its effect in Sec. 4.3.

Inference. At test time the 3D stage passes the fields through robust $\boldsymbol{\Omega}$-triangulation, a Kabsch alignment, and a *single* LM fit of 15 iterations with the feed-forward shape $\hat{\boldsymbol{\beta}}$ and the unit-matched 3D anchor of Sec. 4.3 (LM damping 10^{-3}, PCA prior weight 0.01). Complete pseudocode, including the LM loop, is given in the supplement.

Parameter-free, not training-free. The solver has no learnable weights, yet the model *is* trained through it (Stage B), while at inference the 3D stage is pure optimization.

4 Experiments

4.1 Setup

Datasets and protocol. We evaluate on the five multi-view benchmarks used by POEM-v2 [63]: DexYCB [3] (8v, test), HO3D [12] (5v, test), ARCTIC [6] (8v, val-p1), InterHand2.6M [34] (8v, val), and OakInk [60] (4v, test). We additionally measure *zero-shot* transfer on two datasets outside the training mix,

with no fine-tuning: an internal multi-view hand dataset (4v) and GigaHands (6v). We report on the *full* official evaluation splits, at each dataset's maximum view count and a fixed seed. Splits, view counts, and seed follow POEM-v2's protocol exactly. The metrics are the mean per-vertex and per-joint position error (MPVPE, MPJPE) and their Procrustes-aligned (PA) variants, in mm, as absolute error in the first camera frame. Following POEM-v2, joints are regressed from the predicted mesh for both methods, as the raw joint annotations follow a different joint convention. InterHand2.6M frames contain two interacting hands, while both methods are single-hand estimators (Sec. 3.1), and the crop boxes stored with the released shards are corrupted for a subset of captures (out-of-frame joints inflate them by an order of magnitude). We therefore rebuild each view's crop from the target hand's visible ground-truth joints, *identically for both methods*, so each crop tightly contains the evaluated hand. This *oracle-crop* protocol measures relative accuracy under matched localization, not robustness to a detector. Both models in Table 1 run through our evaluation code in one session under one identical protocol, from the authors' released checkpoints. Our re-evaluation reproduces POEM-v2's published values on DexYCB (7.0 vs 7.00) and HO3D (9.1 vs 9.16). On OakInk it reads 8.8 vs the published 7.91 because the released shards encode a different official split than the published row, so we report the like-for-like value on the released split for both methods. Because both models are evaluated on the complete splits, Table 1 reports population means. Per-scene bootstrap confidence intervals are at most ± 0.09 mm on every dataset and metric (supplement), so all per-dataset margins of 0.5 mm or more lie outside scene-level noise, while the 0.06 mm ARCTIC MPVPE margin does not. These intervals quantify scene-sampling variation. Training replications (a shape-head refit and a full retrain with a new seed) shift the means by at most 0.2 mm (0.5 on InterHand2.6M) and flip no verdict (supplement).

Baselines. The comparison is restricted to methods sharing UA-Fit's MANO output space, i.e., POEM-v2-param (the `medium_MANO` checkpoint, 40.3M head). Non-parametric free-vertex models (POEM-v2 small/large, MvP, PE-MeshTR, FTL-MeshTR) target a different output space, and the free-vertex POEM-v2-large sets the accuracy ceiling (1–3 mm ahead, Sec. 5).

4.2 Main results: parametric parity at a fraction of the size

Table 1 shows the main result. Against POEM-v2's *trained* 40.3M parametric head, UA-Fit (whose 3D solver carries *no* learned parameters) is more accurate on HO3D (-1.0 mm MPVPE) and on InterHand2.6M (-3.2), on par on ARCTIC (7.5 vs 7.6), and less accurate on DexYCB ($+0.9$) and OakInk ($+0.8$). Averaged over the five benchmarks it attains 6% lower MPVPE (8.4 vs 8.9 mm) and 11% lower MPJPE (7.7 vs 8.6 mm), and is more accurate on three of the five MPJPE comparisons. InterHand2.6M is the informative case: direct parameter regression degrades there (12.0 mm MPVPE) while the anchored MANO fit stays close to the robust triangulation (8.8 mm) and MANO-consistent (Fig. 3). The DexYCB and OakInk deficits reflect the 2D evidence rather than the solver: the fit already lies within 2.1–2.5 mm of its own triangulation target, and a roughly $4\times$ longer

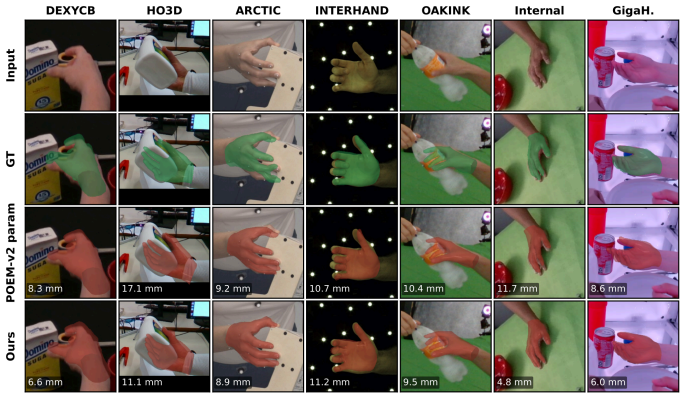


Fig. 3: Qualitative results. One scene per benchmark (columns 1–5) and the two zero-shot sets (columns 6–7): ground-truth MANO (green), POEM-v2-param, and UA-Fit (salmon) reprojected on the target view (MPJPE annotated). Single illustrative scenes, with per-dataset averages in Table 1 (POEM-v2-param is more accurate on average on DexYCB and OakInk, and single scenes can deviate from the averages, as the InterHand column shows).

training schedule reaches 7.3 mm MPVPE on DexYCB, so the gap is training budget, not architecture.

Zero-shot transfer. The advantage holds out of distribution: on two datasets neither model was trained on, UA-Fit cuts MPVPE by a third on the internal set (11.5 vs 16.9) and by 3.1 mm on the full GigaHands test set (13.7 vs 16.8), and leads after Procrustes alignment as well (6.7/7.5 vs 8.4/8.1; Table 1).

Efficiency. UA-Fit and POEM share the HRNet-W40 backbone (57.6M), but its learned reconstruction stage is only 2.0M parameters (1.33M evidence heads + 0.67M neck) against the 40.3M parametric fusion head it replaces (20×, and 46× vs POEM-v2-large's 92.8M vertex head), for a 59.6M total, 1.6× smaller than POEM-v2-param. Inference runs a feed-forward shape head and a single LM solve at ∼60 hands per second (batch 16, one H200), with no test-time shape optimization. Because the batched analytic solver is nearly flat in batch size, throughput keeps rising with the batch while the feed-forward baseline plateaus (supplement). The advantage lies in throughput rather than latency: at a batch size of one the fixed cost of the LM solve makes UA-Fit about 3× slower, and the two reach parity at batch 16. This profile suits multi-hand capture rigs and storage-constrained edge deployment, and a dedicated edge study is left to future work.

4.3 Solver study

This study isolates the solver: every row reuses the same *frozen* backbone, heads, and dense evidence on a fixed evaluation subset and swaps only the 3D stage. All rows are absolute MPVPE, with row a our pipeline matching Table 1, and costs are given per iteration on one GPU.

Table 2: Solver ablation on a fixed frozen 2D front-end and evaluation subset (MPVPE, mm, with relative per-iteration latency). Row a is UA-Fit, rows b–d swap only the 3D stage, isolating the uncertainty weighting (b) and the solver itself (c–d), and all variants add zero parameters.

	evidence → solver	DexY.	HO3D	ARC.	IH	Oak.	cost
(a)	**DOVF energy → block-w. LM (UA-Fit)**	**7.9**	**8.1**	**7.5**	**8.8**	**9.6**	1×
(b)	(a), no learned $\boldsymbol{\Omega}$ (solver term)	8.1	8.2	7.9	8.9	9.7	1×
(c)	(a), exact autograd Jac. (Theseus)	12.3	12.1	10.6	12.2	12.7	14×
(d)	(a), L-BFGS solver (warm)	10.7	12.4	12.1	8.9	14.2	3×

Analytic vs. exact Jacobian, and vs. L-BFGS. With everything else identical (energy, iterations, damping, frozen front-end), the analytic articulated Jacobian agrees with the exact autograd Jacobian to cosine 0.998 (relative Frobenius error 6.5% at all 23k linearizations), at 0.4 vs 5.7 ms per LM iteration, about 14× lower cost. The exact Jacobian, however, converges to a worse *mesh* (Table 2 row c, +3.1–4.4 mm MPVPE): free to use the full skinning, it fits the 2D joints with anatomically implausible finger poses (3–5× the joint-limit violation depth, supplement), whereas the analytic Jacobian stays on the kinematic manifold, making the approximation both faster and more robust. Replacing the damped LM with a general-purpose L-BFGS on the same energy is markedly worse even when initialized identically (row d), and without that initialization it converges to a badly wrong global orientation (20–45 mm). The anisotropic trust region and articulated structure, not just the objective, make the fit reliable.

Weighting ablation. Removing the learned precision $\boldsymbol{\Omega}$ from the solver term at evaluation time, on the same weights (no training confound), degrades every benchmark (Table 2 row b), and removing it from the triangulation as well costs +0.6 to +1.5 mm (+1.0 mean), the triangulation that seeds the fit being the dominant site (supplement). Isotropizing the learned $\boldsymbol{\Omega}$ (magnitude kept, directionality discarded) changes nothing: the operative signal is a well-ranked confidence *magnitude* (σ-vs-error r=0.46–0.69; details in the supplement). Separately, training through the solver (Stage-B) improves on the Stage-A 2D-pretrained fields on every benchmark (MPJPE −1.0 to −2.2 mm, MPVPE up to −2.4 mm), while the 2D fields alone already give a usable fit (supplement). The solver stays *parameter-free* (Table 1).

One refinement completes the solver: rescaling the 3D-anchor weight to the pixel-space energy units improves four of the five benchmarks by 0.03 to 0.76 mm (supplement; the Table 1 setting).

Backward: implicit vs. unrolled differentiation. The LM solve is differentiable either by unrolling every iterate (memory grows with iterations) or by implicit differentiation of the converged fixed point ($O(1)$ memory). The implicit gradient is the exact sensitivity of the converged solution, and unrolling *approaches* it as the forward converges (cosine 0.88–0.96 at convergence on three of four settings; supplement). It is thus not a lossy approximation but the convergence limit of unrolling, at 12–14× cheaper backward (15.6 → 1.2 ms) and flat memory. We train with it.

Table 3: Inference-time anatomical-validity energy on the released model (mm, ↓; protocol and base of Table 1). The barrier is added at evaluation time only (no retraining; w=50, last 5 LM iterations); construction, weight robustness, and paired significance in the supplement.

	MPVPE		MPJPE		PA-MPVPE	
dataset (views)	base	$+E_{\text{anat}}$	base	$+E_{\text{anat}}$	base	$+E_{\text{anat}}$
DexYCB (8v)	7.9	**7.8**	7.5	**7.4**	4.9	4.9
HO3D (5v)	8.1	**8.0**	7.5	**7.4**	5.1	**4.9**
ARCTIC (8v)	7.5	**7.4**	6.6	**6.5**	5.7	**5.6**
InterHand (8v)	8.8	**8.7**	7.8	**7.7**	6.2	**6.1**
OakInk (4v)	9.6	**9.2**	8.9	**8.8**	5.7	**5.1**

Inference-time composability. Because 3D is solved by minimizing an explicit energy, extra objectives compose *at inference* with no retraining: the anatomical-validity barrier E_{anat} (Sec. 3.4) added at test time gives a small but consistent gain on every benchmark, 0.1 to 0.4 mm MPVPE and up to 0.6 mm PA-MPVPE on OakInk (Table 3; construction and paired significance in the supplement).

5 Limitations and Future Work

UA-Fit inherits the limits of its evidence. At two views a depth deficit along the short baseline favors dense learned fusion, so we scope UA-Fit to the ≥3-view regime of real rigs (supplement). It is less accurate on DexYCB and OakInk, where the limit is the 2D evidence rather than the solver, while the ARCTIC margin is within run-to-run noise. Hand localization is assumed given, with crops centered on the target hand for both methods. Finally, the non-parametric POEM-v2-large model remains more accurate, likely needing image-conditioned shape evidence beyond joint-level fields.

Future work. Because the normal equations accept arbitrary precision-weighted terms (Sec. 4.3), external evidence such as a stronger 3D anchor, single-view keypoints (WiLoR [44]), or an ensemble can enter the same solver at test time and be distilled into one MANO-consistent mesh.

6 Conclusion

We showed that calibrated multi-view hand recovery needs no large learned 3D stage: a 2.0M-parameter 2D evidence network with a zero-parameter analytical solver matches POEM-v2's trained parametric head across five benchmarks, with MANO-consistent output. What that head learns can instead come from known geometry and a parametric prior.

Acknowledgment. We gratefully acknowledge the funding of the Lighthouse Initiative Geriatronics by StMWi Bayern (Project X, grant no. 5140951) and LongLeif GaPa GmbH (Project Y, grant no. 5140953).

References

1. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
2. Boukhayma, A., de Bem, R., Torr, P.H.S.: 3D hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1902.03451
3. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: DexYCB: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
4. Chen, X., Liu, Y., Dong, Y., Zhang, X., Ma, C., Xiong, Y., Zhang, Y., Guo, X.: MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), arXiv:2112.02753
5. Chen, X., Liu, Y., Ma, C., Chang, J., Wang, H., Chen, T., Guo, X., Wan, P., Zheng, W.: Camera-space hand mesh recovery via semantic aggregation and adaptive 2D-1D registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), arXiv:2103.02845
6. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
7. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
8. Fu, R., Zhang, D., Jiang, A., Fu, W., Funk, A., Ritchie, D., Sridhar, S.: GigaHands: A massive annotated dataset of bimanual hand activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), arXiv:2412.04244
9. Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., Yuan, J.: 3D hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
10. Gu, K., Yang, L., Yao, A.: Removing the bias of integral pose regression. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
11. Güler, R.A., Neverova, N., Kokkinos, I.: DensePose: Dense human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018), arXiv:1802.00434
12. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: HOnnotate: A method for 3D annotation of hand and object poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
13. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1904.05767
14. He, Y., Yan, R., Fragkiadaki, K., Yu, S.I.: Epipolar transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), arXiv:2005.04551

15. Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), arXiv:1905.05754
16. Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A **32**(5), 922–923 (1976)
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018), arXiv:1712.06584
18. Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems (NeurIPS) (2017), arXiv:1703.04977
19. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), arXiv:1909.12828
20. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1905.03244
21. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic modeling for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021), arXiv:2108.11944
22. Kulon, D., Güler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
23. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: An accurate O(n) solution to the PnP problem. International Journal of Computer Vision (IJCV) **81**(2), 155–166 (2009)
24. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021), arXiv:2107.11291
25. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), arXiv:2011.14672
26. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), arXiv:2203.09364
27. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), arXiv:2012.09760
28. Lin, K., Wang, L., Liu, Z.: Mesh Graphormer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021), arXiv:2104.00272
29. Liu, Y., Long, X.X., Habermann, M., Yang, X., Lin, C., Liu, Y., Ma, Y., Wang, W., Liu, L.: HGGT: Robust and flexible 3D hand mesh reconstruction from uncalibrated images. arXiv preprint arXiv:2603.23997 (2026)
30. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)

31. Moon, G.: Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023), arXiv:2303.13652
32. Moon, G., Lee, K.M.: I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020), arXiv:2008.03713
33. Moon, G., Shiratori, T., Lee, K.M.: DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020), arXiv:2008.08213
34. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
35. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016), arXiv:1603.06937
36. Nibali, A., He, Z., Morgan, S., Prendergast, L.: Numerical coordinate regression with convolutional neural networks. arXiv preprint arXiv:1801.07372 (2018)
37. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: AssemblyHands: Towards egocentric activity understanding via 3D hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023), arXiv:2304.12301
38. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (TMLR) (2024), arXiv:2304.07193
39. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: HandOccNet: Occlusion-robust 3D hand mesh estimation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022), arXiv:2203.14564
40. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1904.05866
41. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3D with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), arXiv:2312.05251
42. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: PVNet: Pixel-wise voting network for 6DoF pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1812.11788
43. Pineda, L., Fan, T., Monge, M., Venkataraman, S., Sodhi, P., Chen, R.T.Q., Ortiz, J., DeTone, D., Wang, A., Anderson, S., Dong, J., Amos, B., Mukadam, M.: Theseus: A library for differentiable nonlinear optimization. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), arXiv:2207.09442
44. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), arXiv:2409.12259

45. Prokudin, S., Gehler, P., Nowozin, S.: Deep directional statistics: Pose estimation with uncertainty quantification. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018), arXiv:1805.03430
46. Qiu, H., Wang, C., Wang, J., Wang, N., Zeng, W.: Cross view fusion for 3D human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), arXiv:1909.01203
47. Remelli, E., Han, S., Honari, S., Fua, P., Wang, R.Y.: Lightweight multi-view 3D pose estimation through camera-disentangled representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), arXiv:2004.02186
48. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (Proc. SIGGRAPH Asia) **36**(6) (2017)
49. Song, J., Chen, X., Hilliges, O.: Human body model fitting by learned gradient descent. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020), arXiv:2008.08474
50. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1902.09212
51. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018), arXiv:1711.08229
52. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020), arXiv:2008.11200
53. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment — a modern synthesis. In: Vision Algorithms: Theory and Practice, pp. 298–372. Springer (2000)
54. Tu, H., Wang, C., Zeng, W.: VoxelPose: Towards multi-camera 3D human pose estimation in wild environment. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020), arXiv:2004.06239
55. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), arXiv:2503.11651
56. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024), arXiv:2312.14132
57. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018), arXiv:1804.06208
58. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing Systems (NeurIPS) (2022), arXiv:2204.12484
59. Yang, J., Li, J., Li, G., Shen, Z., Wu, H.Y., Fan, Z., Huang, H.: MLPHand: Real time multi-view 3D hand reconstruction via MLP modeling. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024), arXiv:2406.16137
60. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: OakInk: A large-scale knowledge repository for understanding hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

61. Yang, L., Xu, J., Zhong, L., Zhan, X., Wang, Z., Wu, K., Lu, C.: POEM: Reconstructing hand in a point embedded multi-view stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023), arXiv:2304.04038
62. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: CPF: Learning a contact potential field to model the hand-object interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
63. Yang, L., Zhong, L., Zhu, P., Zhan, X., Kong, J., Xu, J., Lu, C.: Multi-view hand reconstruction with a point-embedded transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2025), arXiv:2408.10581
64. Zanfir, A., Zanfir, M., Bazavan, E.G., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: Neural descent for visual 3D human pose and shape. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021), arXiv:2008.06910
65. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), arXiv:1910.06278
66. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: PyMAF: 3D human pose and shape regression with pyramidal mesh alignment feedback loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021), arXiv:2103.16507
67. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular RGB image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), arXiv:1902.09305
68. Zhang, Z., Wang, C., Qiu, W., Qin, W., Zeng, W.: AdaFuse: Adaptive multiview fusion for accurate human pose estimation in the wild. International Journal of Computer Vision (IJCV) **129**(3), 703–718 (2021), arXiv:2010.13302
69. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019), arXiv:1812.07035
70. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single RGB images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017), arXiv:1705.01389
71. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: 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 (ICCV) (2019)