2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust Hand Tracking from Visual-Inertial Fusion

Hyelim Choi[1], Hyunreal Park[1], Harim Ji[1], Somang Lee[1], Youngseon Lee[1], Yongseok Lee[2], and Dongjun Lee[1]

***Abstract*— Hand tracking plays a key role in capturing and transferring dexterous human manipulation skills to robots. However, achieving reliable tracking across diverse conditions and during complex interactions (e.g., object manipulation) remains challenging. A promising solution is to combine wearable sensors such as IMUs with vision, where previous studies have handled the vision input by attaching markers to wearables or by relying on depth data to avoid the domain gap in color images. In this work, we present a hand tracking framework that fuses inertial measurements with state-of-the-art vision methods, eliminating the need for markers while fully exploiting visual cues. For this, we introduce a dataset generation scheme that produces synthetic and real data for the target glove using a compact setup, without manual annotation. Using the dataset, we train the keypoint detection network that predicts the likelihood of an image for keypoints, designed based on a lightweight vision transformer (ViT) for real-time usage. Based on the network prediction, the IMU-propagated pose is used as a prior in probabilistic inference to estimate the keypoint positions and uncertainties. Tracking primarily relies on high-rate IMU updates for fast motion estimation, while the pose is corrected through factor graph optimization. The framework is validated in challenging scenarios, demonstrating its robustness and accuracy, and can be used for high-quality demonstration data acquisition and teleoperation for dexterous manipulation.**

I. INTRODUCTION

Despite its significance across diverse fields, hand tracking remains challenging due to the inherent limitations of sensor modalities. For instance, IMUs suffer from drift, soft sensors fail under contact and lack durability, and magnetic sensors are vulnerable to electromagnetic interference. More recently, vision-based methods have achieved impressive results, but they remain susceptible to occlusion, particularly during object manipulation. Combining vision with wearable sensors (e.g., IMUs) offers a promising solution to overcome occlusion, where prior studies have processed the vision input with visual markers or relied solely on depth to avoid the domain gap inherent in RGB-based methods. While state-of-the-art vision techniques can benefit the framework by eliminating the need for markers and fully exploiting rich visual information, this direction remains underexplored. Leveraging such vision techniques, however, requires a specialized dataset for target wearables, whose generation often demands an exhaustive and costly annotation process.

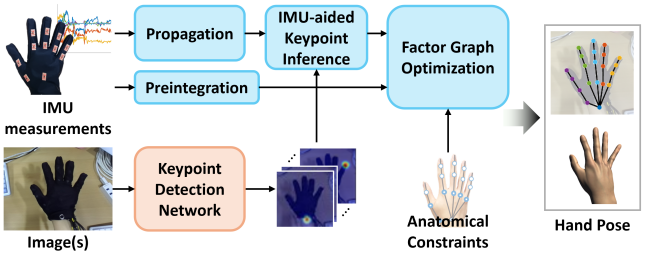

Fig. 1. Framework of the proposed visual-inertial hand tracking.

In this paper, we present a hand tracking framework that combines modern vision methods with inertial data from glove-mounted IMUs (Fig. 1). To train the glove-specific keypoint detection network, we propose an efficient synthetic and real dataset generation scheme that eliminates the need for multi-view studio setups and manual annotations. The network is designed to be lightweight for real-time usage and is trained to predict the likelihood of the input RGB image conditioned on the keypoint locations. Then, with the aid of the prior from the IMU-propagated pose, the 2D keypoint positions and their uncertainties are estimated. Tracking is mainly conducted through IMU updates, which allows high tracking speed. The hand pose is corrected via factor graph optimization, which integrates the predicted keypoint positions, preintegrated IMU signals, and anatomical constraints. The proposed framework is validated in challenging scenarios, showing its robustness and accuracy compared to existing methods.

By providing reliable and precise hand tracking, our method can be directly used for high-quality demonstration data acquisition and teleoperation essential for dexterous manipulation tasks. Furthermore, the acquired dataset can be enriched with hand-object contact information, either by calculating contact forces in simulation or by attaching force sensors (e.g., force sensitive resistors) to the glove in the real world. Haptic feedback can also be added via actuators (e.g., linear resonant actuators) on the glove, enhancing teleoperation performance.

Our contribution is summarized as follows:

1) We introduce an efficient dataset generation scheme for training the glove-specific keypoint detection network.
2) We design a keypoint detection network based on a lightweight vision transformer (ViT) for real-time use and propose an IMU-aided keypoint inference for robust estimation of keypoint positions and uncertainties.
3) We refine the pose via a factor graph optimization framework for robust and accurate tracking that can be reliably used in various applications, including robotics.

This work was supported by Korea Planning & Evaluation Institute of Industrial Technology (KEIT) grant funded by Ministry of Trade, Industry and Resources (MOTIR) (No. RS-2024-00441872) and Korea Institute of Marine Science & Technology Promotion (KIMST) grant funded by Ministry of Oceans and Fisheries (MOF) (No. RS-2025-02305446).

[1]The authors are with the Department of Mechanical Engineering, IAMD and IOER, Seoul National University, Seoul 08826, South Korea (e-mail: {helmchoi, hyun3655, jiharim0911, hopelee, yslee1765, djlee}@snu.ac.kr).

[2]Yongseok Lee is with the Department of Robotics and Mechatronics, DGIST, Daegu 42988, South Korea (e-mail: yslee@dgist.ac.kr).

Dongjun Lee and Yongseok Lee are co-corresponding authors.

II. Related Works

A. Vision

Monocular vision is extensively studied recently, where neural networks were trained to estimate either the hand skeleton [1], [2] or mesh [3]–[7]. To overcome occlusion, some leveraged image sequences [5], and others designed network structures to exploit the non-occluded information to better estimate the occluded region [4]. However, these works rely on inductive bias (e.g., anatomy, object interaction, and temporal smoothness) to estimate the most likely hand pose rather than actually measuring it, thus frequently failing to deal with severe occlusion. In addition, recent large-scale networks trained with a vast amount of data are computationally expensive, which hinders real-time usage [7].

Multi-view images from calibrated cameras alleviate the occlusion issue and provide depth information. In [8], the 3D hand skeleton was estimated from five images via optimization. An end-to-end approach was adopted to predict either the hand mesh or skeleton from a stereo image pair [9], multi-view images [10], and four wide-FoV monochrome cameras [11]. Although the use of multi-view images alleviates the occlusion problem, dealing with a severe occlusion (e.g., object interaction) is still fundamentally unresolved. Additionally, the multi-view camera setting is bulky and requires careful calibration.

Using depth measurements, the hand pose was obtained via an optimization to match the rendered and measured depths in [12]. More recent works [13], [14] leveraged learning-based approaches to estimate the hand pose. Still, depth has a limited sensor range, is more costly than RGB images, and lacks color, which is a valuable key for hand estimation. An RGB image was exploited with depth in [15]; nevertheless, there remains the fundamental issue of occlusion that is inherently present in vision-based methods.

B. Inertial measurements

Inertial measurements (i.e., acceleration and angular velocity) are utilized by embedding IMUs on a wearable device. Unlike vision, such sensing offers an advantage in that it is unaffected by occlusion. Since using only the inertial measurement leads to drift, magnetometers are typically employed to correct orientation errors [16]–[19]. However, relying on magnetometers compromises the robustness to electromagnetic interference, which can be induced by nearby magnetic objects or electronic devices. Furthermore, positional drift remains unavoidable, and acceleration measurements are susceptible to disturbances caused by contact.

C. Multi-modal fusion

Multiple sensor modalities can be fused to overcome the fundamental limitations of each single modality. In [20], gyroscope and electromyography (EMG) measurements were adaptively integrated through a neural network via the attention mechanism, but it could not track the global hand position and had limited tracking accuracy. A gyroscope was used in [21] with a depth camera to overcome the distortion of depth due to motion blur, but the use of depth still confined the user's motion range and robustness to occlusion. In [22], the global hand pose was tracked using a depth image and IMUs on the arm, while the local hand pose was estimated by loosely coupling independent pose estimates from the depth image and a soft sensor glove. Another solution was proposed in [23], where a stereo camera estimated the positions of colored markers on the glove, which are complementarily fused with inertial measurements. However, to the best of our knowledge, the combination of state-of-the-art vision techniques and inertial measurements for hand tracking has been rarely explored. A recent study [24] employed attention-based fusion, but overlooked IMU dynamics and multi-modal uncertainties critical for robust tracking. Moreover, the computational complexity of the attention mechanism may challenge real-time feasibility. While similar approaches have been adopted in human pose estimation [25]–[27], hand tracking poses distinct challenges. Specifically, wearable-induced visual variations are more pronounced for hands than for the body, thereby limiting the utility of off-the-shelf estimators and datasets. Furthermore, these works rely on global orientation from IMUs rather than angular velocity, requiring high-precision IMUs equipped with magnetometers that are susceptible to electromagnetic disturbances.

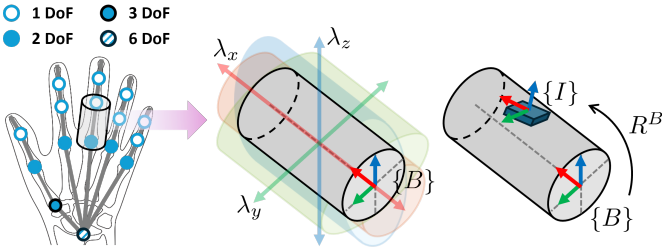

Fig. 2. The hand skeletal model and kinematic parameters of a segment: λ and R^B. The hand segment and IMU coordinate systems are denoted $\{B\}$ and $\{I\}$, respectively.

III. Preliminaries

A. Hand model

Following [23], we model the hand as a 26-degrees-of-freedom (DoF) skeleton comprised of 16 segments and 16 joints, as shown in Fig. 2. The wrist has 6 DoF (i.e., position and orientation), and each finger has three segments and three joints. Except for the thumb, the fingers have one MCP joint (2 DoF), one PIP joint (1 DoF), and one DIP joint (1 DoF). The thumb is more flexible: we model it with one CMC joint (3 DoF), one MCP joint (2 DoF), and one IP joint (1 DoF).

Two kinematic parameters are assigned for each IMU-attached segment: the scale parameter $\lambda \in \mathbb{R}^3$ and the IMU misalignment $R^B \in SO(3)$ (see Fig. 2). The scale parameter represents the size of the segment relative to the pre-defined nominal size, and the IMU misalignment is the orientation of the attached IMU relative to the segment coordinate system.

B. Glove design

We fabricate the outer layer of the glove in a solid color, which simplifies the manufacturing process and yields a plain design that is less conspicuous in everyday use. The glove is equipped with 11 IMUs: one on the dorsum and two on each finger (proximal and intermediate phalanges).

The sensors are adhered to the inner layer using hot-melt adhesive and placed approximately at the nominal centers of the phalanges. Precise placement is not required, as sensor position and orientation offsets are identified through kinematic calibration. We omit IMUs on the distal phalanges and estimate their poses via a hand synergy model, assuming the distal joint angles are proportional to the proximal ones. While this may not capture independent distal joint movements during passive interactions (e.g., contact), it remains practically sufficient for most tasks. Note that the system is readily scalable to 16 IMUs to achieve full tracking, obviating this synergy-based assumption. We use LSM6DS33 6-axis IMUs and a Teensy 4.0 MCU to process the IMU signals, which is connected to the PC via USB. The glove weighs 58~60 [g] in total, depending on its size.

IV. Keypoint Detection Network

A. Dataset generation

First, we generate a synthetic dataset by rendering images and ground truth (GT) keypoints (i.e., joints and fingertips) from specified joint angles using a black hand in a photorealistic rendering tool [28]. The joint angles, randomly sampled from Gaussian distributions and clipped to pre-defined limits, are commanded to a virtual proxy in a physics simulation [29] to resolve inter-segment collisions. Then, the images are augmented by filling the background with images in an open-source dataset [30] and placing random circular/rectangular patches of random color/black/random noise that occlude the hand. Each image is augmented to five images (total 100000 images), and the samples are displayed in Fig. 3.

To bridge the synthetic-to-real gap, we further generate a real dataset. We exploit tracking results of VIST [23] as the GT, which provides accurate tracking under a compact setting (i.e., only one stereo camera). While tracking, we save the images used for tracking along with the 3D joint and fingertip positions. Then, we remove the markers from the images using an inpainting algorithm [31], as done in [32]. We detect the markers on the image using the HSV criterion and annotate markers that are missed due to their viewing angles to create masks for inpainting. Each marker-free inpainted image is augmented to 12 images similar to the synthetic images, and additionally by random brightening and darkening (see Fig. 3), resulting in a total of 12036 images. This semi-automatic dataset generation scheme neither requires a laborious manual annotation nor a controlled multi-view studio with tens of cameras.

B. Network configuration

To ensure real-time feasibility, our network is built upon FastViT-MA36, a variant of the lightweight vision transformer architecture [33]. The network consists of a stem network, followed by two stages composed of 4 RepMixer token mixers and a patch embedding layer positioned between them. After batch normalization, the output of the last stage is upsampled and added to the batch-normalized output of the first stage before being passed to the subsequent convolution layers. The input image is cropped to the region of interest (RoI) of the hand and resized to 256×256 before being passed to the network. The RoI is easily computed from the current pose estimate thanks to our real-time tracking framework, and its size is set proportional to the uncertainty of the estimation. The network outputs a set of 64×64 heatmaps, each corresponding to a keypoint and representing the likelihood of the image given the 2D keypoint position. This heatmap representation enables both keypoint position estimation and uncertainty calculation (e.g., a uniform heatmap implies high uncertainty), which is crucial for fusing information from multiple sensor modalities.

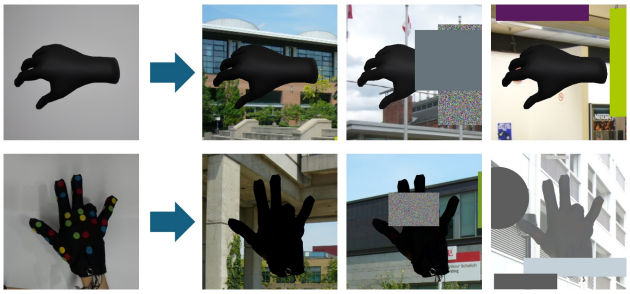

Fig. 3. Samples of augmented synthetic (top) and real (bottom) images. Both datasets are augmented with randomized backgrounds and occlusion patches to enhance robustness. The real images are further augmented by brightness/darkness variations after marker inpainting.

C. Training

For the network training, we generate GT heatmaps from the 3D keypoint positions of the dataset. The GT heatmaps are 2D Gaussians with a predefined standard deviation, centered at the projected keypoint positions, and scaled to a maximum of 1. Although alternative loss functions, such as soft-argmax, can be adopted, we found that they hinder training convergence and lead to suboptimal performance.

We first train the network on the synthetic dataset for 250 epochs and further train for 500 more epochs using the real dataset. As done in [34], the images and GT heatmaps are rotated, translated, and scaled randomly during training to improve generalization. We use AdamW optimizer with a learning rate of 0.0005 to train the network.

V. Visual-Inertial Tracking

A. Formulation

The hand is represented by a state variable, which is the concatenation of kinematic parameters and IMU state variables

$$\mathcal{X}(t) = [\mathbf{x}_1(t), \lambda_1, R_1^B, \mathbf{x}_2(t), \lambda_2, R_2^B, \dots, \mathbf{x}_N(t), \lambda_N, R_N^B]$$

where t is the timestamp, N is the number of IMUs, $\mathbf{x}_i$ is the state of the i-th IMU, and $\lambda_i \in \mathbb{R}^3$ and $R_i^B \in SO(3)$ are the kinematic parameters of the segment on which the i-th IMU is attached. The state of the i-th IMU is defined as

$$\mathbf{x}_i(t) = [p_i(t), v_i(t), R_i(t), b_{a,i}(t), b_{g,i}(t)]$$

where $p_i \in \mathbb{R}^3$ is the global position, $v_i \in \mathbb{R}^3$ is the global velocity, $R_i \in SO(3)$ is the global orientation, $b_{a,i} \in \mathbb{R}^3$ is the accelerometer bias, and $b_{g,i} \in \mathbb{R}^3$ is the gyroscope bias. The biases are known to be slow-varying, thus we regard the biases as static variables $b_{a,i}(t) \equiv b_{a,i}$, $b_{g,i}(t) \equiv b_{g,i}$, $\forall t$.

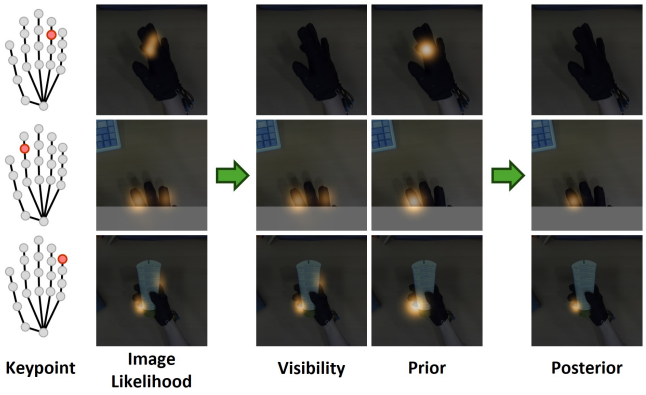

Fig. 4. Probabilities of a keypoint represented in heatmaps (from left to right): the image likelihood $p(I|u,\delta=1)$, visibility $\sum_\delta p(I|u,\delta)\cdot p(\delta|\hat{\boldsymbol{x}})$, unnormalized 2D prior $p(u;\pi(\hat{\boldsymbol{x}}))$, and the unnormalized posterior probability $p(u,I)$. The corresponding keypoint is marked red on the left, and black/white colors indicate low/high probability values, respectively.

Our framework operates with either a monocular camera or multiple synchronized cameras (e.g., stereo), and we assume that the camera poses ($p_{GC}(t)\in\mathbb{R}^3$ and $R_{GC}(t)\in SO(3)$) are known. The camera poses can be obtained with localization algorithms that are frequently embedded in commercial cameras and head-mounted displays (HMDs). The hand pose is primarily updated with every IMU reading via IMU propagation, where the IMU state at time t is propagated using the well-known model [35]. Upon image acquisition, the corresponding timestamp t' is marked as a keyframe timestamp, and the acquired image(s) are processed by the keypoint detection network (Sec. IV) and used in the IMU-aided keypoint inference (Sec. V-B). Then the inference results at keyframe timestamps and IMU signals are used to optimize the hand poses at keyframe timestamps, correcting the IMU-propagated pose estimates (Sec. V-C).

B. IMU-aided keypoint inference

Although the keypoint detection network can independently learn to predict 2D keypoint positions and uncertainties given an image, some challenging scenarios introduce pose ambiguities that are hard to resolve solely from the image. Many vision-only hand pose estimation methods take advantage of inductive bias granted in the large-scale dataset, yet they still fail in out-of-dataset scenarios (e.g., object, environment, and postures) and under severe occlusion. Instead of relying on the inductive bias, we leverage the latest hand pose estimate from the IMU signals, which easily resolves the pose ambiguity present in the image. Since the pose estimate is updated through high-frequency IMU updates, it provides an up-to-date prediction that is close to the actual pose even in challenging scenarios (e.g., fast motion). From the propagated state, we predict the 3D keypoint positions $\hat{\boldsymbol{x}}=[\hat{x}_1,\hat{x}_2,\dots,\hat{x}_K]$ in the camera frame, which we take as a prior for the inference, as $\hat{x}_k = R_{GC}^{-1}\left(\hat{p}_{i_k}+\hat{R}_{i_k}(R_{i_k}^B)^{-1}\mathrm{diag}(\lambda_{i_k})L_{i_k,k}-p_{GC}\right)$, where i_k is the index of the IMU that is attached to the same hand segment with the keypoint, and $L_{i_k,k}\in\mathbb{R}^3$ is the position of the keypoint with respect to the segment coordinate system centered at the IMU position.

The inference can be formulated as an inference on the conditional probability of the 2D keypoint positions $\boldsymbol{u}=[u_1,u_2,\dots,u_K]$ given an image. We assume that the keypoints are conditionally independent (i.e., $p(\boldsymbol{u}|I)=\prod_k p(u_k|I)$), thus we describe the formulation for each k-th keypoint and omit the subscript k for the rest of the subsections unless necessary. The propagated 3D keypoints $\hat{\boldsymbol{x}}$ are integrated as a prior in the conditional probability as

$$p(u|I)\propto p(u,I)=\int_{\boldsymbol{x}}\sum_{\delta} p(I|u,\delta)\cdot p(\delta|\boldsymbol{x})\cdot p(u|\boldsymbol{x})\cdot p(\boldsymbol{x}) \tag{1}$$

where I is the observed image and $\delta\in\{0,1\}$ is the keypoint visibility indicator (i.e., $\delta=1$ if the keypoint is visible). Each probability term is detailed as follows.

1) Image likelihood $p(I|u,\delta)$: It is given as

$$p(I|u,\delta)=\begin{cases}\mathcal{H}(h(u)) & \text{if } \delta=1\\ \epsilon & \text{if } \delta=0\end{cases} \tag{2}$$

where $\mathcal{H}$ is the corresponding heatmap predicted by the network, $h(\cdot)$ is a function mapping a 2D keypoint coordinate on the image to the heatmap coordinate, and ϵ is a constant.

2) Visibility probability $p(\delta|\boldsymbol{x})$: It represents the probability of keypoint visibility, accounting for self-occlusion and camera FoV. Self-occlusion is checked using a simplified hand model comprised of primitives: a box (palm) and cylinders (finger segments). The primitive sizes are obtained from the kinematic parameters calculated through the calibration. For each camera ray pointing to a 3D keypoint, we calculate the margin of intersection with all primitives. Then, we design a metric $p(\delta_{\text{occ}}|\boldsymbol{x})$ using a sigmoid function that returns a value close to 1 when an intersection occurs and 0 otherwise. Similarly, for the FoV, we design a metric $p(\delta_{\text{fov}}|\boldsymbol{x})$ that is close to 1 when the keypoint lies inside of the FoV and 0 otherwise. Multiplying all these metrics, we obtain $p(\delta|\boldsymbol{x})=p(\delta_{\text{occ}}|\boldsymbol{x})\cdot p(\delta_{\text{fov}}|\boldsymbol{x})$.

3) Camera projection $p(u|\boldsymbol{x})$: This is from camera projection of the corresponding 3D keypoint $\boldsymbol{x}_k$ and can be represented as a Dirac delta function $\delta(u-\pi(\boldsymbol{x}_k))$.

4) 3D Prior $p(\boldsymbol{x})$: This prior on the 3D keypoint position is modeled as a Gaussian distribution, with its mean at the corresponding latest IMU-propagated estimate $\hat{\boldsymbol{x}}_k$, i.e., $\mathcal{N}(\hat{\boldsymbol{x}}_k,\Sigma)$. The covariance $\Sigma\in\mathbb{R}^{3\times3}$ is obtained from the optimization during tracking.

Since $p(x)$ in Eq. 1 is small outside the neighborhood of $\boldsymbol{x}=\hat{\boldsymbol{x}}$, we assume that the effect of $p(\delta|\boldsymbol{x})$ on $p(u|I)$ outside this region is negligible. Thus, we approximate it by a nominal value, $p(\delta|\boldsymbol{x})\approx p(\delta|\hat{\boldsymbol{x}}),\ \forall\boldsymbol{x}$, which makes the formulation tractable. Then Eq. 1 can be rewritten as

$$\begin{aligned} p(u|I)&\propto p(u,I)\\ &\approx\left\{\sum_{\delta}p(I|u,\delta)\cdot p(\delta|\hat{\boldsymbol{x}})\right\}\cdot\int_{\boldsymbol{x}}p(\boldsymbol{x})p(u|\boldsymbol{x})\\ &=\left\{\sum_{\delta}p(I|u,\delta)\cdot p(\delta|\hat{\boldsymbol{x}})\right\}\cdot p(u;\pi(\hat{\boldsymbol{x}})) \end{aligned} \tag{3}$$

where $p(u;\pi(\hat{\boldsymbol{x}}))$ is the 2D Gaussian probability distribution projected on the image plane from the 3D Gaussian $p(\hat{\boldsymbol{x}})$.

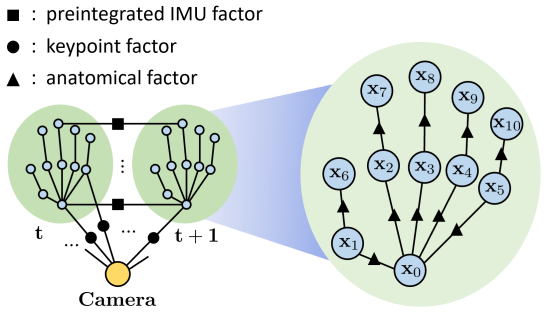

Fig. 5. Factor graph for optimizing the hand pose. Preintegrated IMU factors, keypoint factors, and anatomical factors are integrated.

The 2D Gaussian has its mean on $\pi(\hat{\boldsymbol{x}})$ and its covariance is $JR_{GC}^{T}\Sigma R_{GC}J^{T}$, where $J \in \mathbb{R}^{2\times 3}$ is the jacobian of the projection. In practice, we scale up Σ to $\kappa\Sigma$ ($\kappa > 1$) to give more weight to the vision estimation (e.g., $\kappa = 2.0$ in the experiments in Sec. VI). Fig. 4 shows examples of the probabilities in Eq. 3, where the likelihood predicted by the keypoint network is ignored for a self-occluded keypoint (row 1) and the prior biases the ambiguous likelihood toward the correct position (rows 2 and 3). Note that when inferred keypoints are uncertain or missing for some time, the prior may be misleading due to estimation errors, but its increased uncertainty reduces its influence on the inference.

For an efficient sensor fusion, we approximate $p(u|I)$ as an isotropic Gaussian probability distribution and obtain its mean $\mu \in \mathbb{R}^2$ and covariance $\sigma \cdot I_{2\times 2}$. Then the joint probability $p(u, I)$ computed from Eq. 3 can be written as

$$p(u, I) = s \cdot p(u|I) \approx \frac{s}{2\pi\sigma} \exp\left(-\frac{1}{2\sigma^2}\|u-\mu\|^2\right) \quad (4)$$

where $s = p(I)$. Using this relation, we obtain σ by matching the area within three-sigma, which is determined by the number of 2D pixel coordinates $q = [i, j]$ for which $p(q, I)$ exceeds the threshold corresponding to three-sigma as

$$\sum_{q}\left[p(q, I) > \frac{s}{2\pi\sigma}\exp\left(-(3\sigma)^2/2\sigma^2\right)\right] = \pi(3\sigma)^2 \quad (5)$$

where $\frac{s}{2\pi\sigma} = \max_q p(q, I)$. The mean μ corresponds to the pixel coordinate where $p(q, I)$ has its maximum value (i.e., argmax), but is computed with soft-argmax to promote smoothness and to handle the low resolution of the heatmap.

C. Pose optimization

We refine the hand pose through factor graph optimization over the state variables, as depicted in Fig. 5. Three factors are integrated: preintegrated IMU factors, keypoint factors, and anatomical factors, as

$$\{\mathcal{X}(t')\}_{t'} = \underset{\{\mathcal{X}(t')\}_{t'}}{\operatorname{argmin}} \sum_{t'} \left\{ \sum_{i=1}^{N} \left\|\mathbf{r}_I^i(t')\right\|^2_{\mathbf{P}_I^i(t')} + \sum_{j=1}^{M}\sum_{k=1}^{K} \left\|\mathbf{r}_V^{j,k}(t')\right\|^2_{\mathbf{P}_V^{j,k}(t')} + \sum_{m\in\mathcal{C}} \|\mathbf{r}_C^m(t')\|^2_{\mathbf{P}_C^m(t')} \right\} \quad (6)$$

where t' is the keyframe timestamp, M is the number of cameras, $\mathbf{r}_I^i$ is the preintegrated i-th IMU residual, $\mathbf{r}_V^{j,k}$ is the k-th 2D keypoint position residual for the j-th camera, $\mathbf{r}_C^m$ is the m-th anatomical constraint residual, m is the constraint index, $\mathcal{C}$ is the set of anatomical constraints, and $\mathbf{P}_I^i, \mathbf{P}_V^{j,k}$, and $\mathbf{P}_C^m$ are the corresponding covariance matrices. The covariance of preintegrated IMU ($\mathbf{P}_I$) is obtained from sensor noise specifications, and that of keypoint position ($\mathbf{P}_V$) is calculated from the uncertainty predicted by the keypoint inference (σ in Eq. 4). The covariance of anatomical constraints ($\mathbf{P}_C$) is an empirically tuned constant. The IMU preintegration follows [36], and the anatomical constraints consist of positional and rotational constraints of hand segments given the kinematic parameters. The constraints for each joint connecting two hand segments are defined as

$$\mathbf{r}_{\mathcal{C}}(m, \mathcal{X}) = \begin{cases} p_p + R_p(R_p^B)^{-1}\mathrm{diag}(\lambda_p)L_{p,J} \\ \quad -(p_c + R_c(R_c^B)^{-1}\mathrm{diag}(\lambda_c)L_{c,J}) \\ \qquad\qquad \cdots \text{if } m \text{ is positional} \\ g(\mathrm{ypr}((R_pR_p^B)^{-1}(R_cR_c^B)), [lb, ub]) \\ \qquad\qquad \cdots \text{if } m \text{ is rotational} \end{cases} \quad (7)$$

where $\star_p$ and $\star_c$ denote parent and child segment variables connected by the joint, respectively, and $L_{p,J}$ and $L_{c,J}$ are the joint positions expressed in the segment coordinate systems centered at the corresponding IMU positions. The function g is an element-wise dead zone function with joint angle bounds lb and ub, and $\mathrm{ypr}(\cdot)$ returns roll, pitch, and yaw angles. For real-time optimization, we maintain a fixed window and marginalize variables outside the window.

D. Kinematic calibration

The kinematic parameters (i.e., scale parameters $\{\lambda_i\}$ and IMU misalignments $\{R_i^B\}$) vary according to each user and might change each time the user wears the glove. To account for these variations and enhance tracking performance, these parameters are self-calibrated via offline batch optimization prior to the tracking session. This optimization minimizes a residual augmented with parameter priors in addition to the tracking residual (Eq. 6). For calibration, image and sensor data are collected while the user performs hand motions. Any motion that renders the parameters observable is sufficient; for instance, our evaluation utilizes a sequence (under 10 seconds) featuring hand flipping, translation, and finger articulation. This process does not require GT from motion capture systems, as it leverages the spatial consistency between the two modalities during hand movements.

VI. Evaluation

A. Quantitative evaluation

For the quantitative evaluation, we use a VICON motion capture system with six infrared (IR) cameras and six reflective markers attached to the glove (Fig. 6). The markers are placed on the dorsum and five fingertips, as these are important keypoints and suitable locations for attachment. The VICON system and the hand tracking camera are calibrated in advance. Since method-specific dataset requirements preclude direct comparison on a common dataset, we

Fig. 6. Motion capture setup for quantitative evaluation. We use six calibrated IR cameras (left figure; highlighted in green) and six reflective markers attached to the glove (right figure; marked in red).

assess accuracy for representative poses. We evaluate on free-hand poses exhibiting self-occlusions, as the VICON system frequently loses marker tracking during object interactions.

We recruit ten participants (six males and four females, aged 20–30s) and ask them to mimic displayed poses. Their hands vary in size and shape, with the representative lengths (from the ulnar styloid process to the tip of the middle finger in rest pose) ranging from $17.2 - 21.5\,[\mathrm{cm}]$. Ethics approval was not required for the study, as it involved no physical intervention or identifiable data collection, but informed consent was obtained from all participants. The accuracy is reported as 3D mean per keypoint position error (MPKPE) and percentage of correct keypoints (PCK), both with and without Procrustes alignment (PA), following previous research. The position error is evaluated using offsets that are computed to minimize the mean error between predicted and actual VICON marker positions.

To evaluate the effectiveness of our proposed algorithm and visual-inertial framework, we compare our system in both monocular and stereo configurations with state-of-the-art vision-based methods (FrankMocap [3], H2ONet [5], SimpleHand [6], and HaMeR [7]) and a visual-inertial method (VIST [23]). To ensure a fair comparison, we provide the vision-based models with bare-hand images, along with manually annotated bounding boxes and GT hand scale. All models use publicly available pretrained weights and are evaluated on monocular images, except for H2ONet, which is given three sequential images. Our method and VIST perform real-time tracking, whereas the other methods estimate poses from pre-recorded images. All experiments are conducted on a laptop with an AMD Ryzen 9 7945HX CPU and an NVIDIA GeForce RTX 4080 Laptop GPU.

The accuracy and speed of the algorithms are reported in Tab. I, where our proposed framework generally outperforms other methods. The vision-based algorithms frequently fail to accurately estimate the pose when there is a severe occlusion, thus yielding large position errors. VIST, which uses a stereo camera, shows improved performance, comparable to our stereo-based counterpart. Our framework outperforms VIST under varying lighting conditions, maintaining robust performance, whereas VIST struggles with such changes, as shown in the reduced participant study with three users (Tab. II). This is because the marker detection of VIST should be tuned for each lighting condition for the best performance, while our keypoint detection network is trained under diverse lighting conditions and learns to see the entire image to leverage hand geometry, resulting in improved robustness to lighting variations. Although the tracking speed

TABLE I

QUANTITATIVE COMPARISON RESULTS. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD, AND THE SECOND-BEST IS UNDERLINED. MPKPE AND PA-MPKPE ARE EXPRESSED IN MILLIMETERS.

Method	MPKPE	PCK @15mm	PA-MPKPE	PA-PCK @15mm	Speed [Hz]
FrankMocap	32.41	0.15	17.33	0.54	137
H2ONet	53.95	0.05	24.53	0.28	61
SimpleHand	39.59	0.12	18.94	0.48	72
HaMeR	22.49	0.37	13.01	0.72	41
VIST	<u>12.17</u>	<u>0.74</u>	8.58	0.89	**249**
Ours (mono)	15.29	0.57	<u>7.96</u>	<u>0.94</u>	<u>206</u>
Ours (stereo)	**12.07**	**0.75**	**7.27**	**0.95**	205

TABLE II

PERFORMANCE UNDER VARIOUS LIGHTING CONDITIONS, SIMULATED VIA CAMERA SETTINGS AND QUANTIFIED BY NORMALIZED MEAN IMAGE INTENSITY. FOR EACH METHOD, THE FIRST AND SECOND ROWS PRESENT 3D MPKPE AND 3D PA-MPKPE, RESPECTIVELY.

Method	Mean Image Intensity					Std. Dev.
	0.14	0.18	0.24	0.45	0.92	
VIST	57.05	40.47	17.67	10.60	10.97	20.60
	21.34	18.06	11.64	7.58	7.65	6.22
Ours (stereo)	15.95	13.26	12.66	12.74	13.83	**1.35**
	8.88	8.11	7.90	7.58	7.82	**0.50**

of the vision-based methods is calculated solely based on inference time, VIST and our framework offer significantly faster tracking due to the high sensor rate of IMUs, while our framework also benefits from the lightweight design of the keypoint detection network.

B. Qualitative evaluation

We present qualitative evaluation compared to vision-based [3], [6], [7] state-of-the-art methods in Fig. 7. For a fair comparison with the vision-based methods, our framework is evaluated using monocular images only. The results show that our method estimates hand poses accurately even under self-occlusion and object interaction, whereas other methods struggle, particularly with the occluded parts.

For the visual-inertial state-of-the-art, VIST [23], we showcase a scenario where the marker detection is tuned for a certain ambient lighting before changing illumination. When the lighting condition is changed, few markers are detected, increasing the risk of drift and incorrect fusion of marker detections and IMU data. The first two rows in Fig. 8 show the results before (1st row) and 5 minutes after (2nd row) the light change. Five minutes after the illumination change, drift occurs in VIST where the markers are no longer detected, ruining the IMU-marker correspondence and degrading the estimation, while our method remains robust, as most keypoints are still detected. The bottom two rows depict object interaction scenarios in which most of the markers are obscured by an object. In the third row, the VIST estimation has drifted due to the lack of detected markers on the fingers, while our algorithm still detects keypoints on the fingers (e.g., fingertips), preventing drift. In the last row, the IMU-

Fig. 7. Qualitative evaluation compared to vision-based state-of-the-art methods. Each image column shows the input view (left) and a reference side view (right). The 3D hand meshes are visualized in two views: camera view (left) and side view rotated 90 [deg] to the left (right).

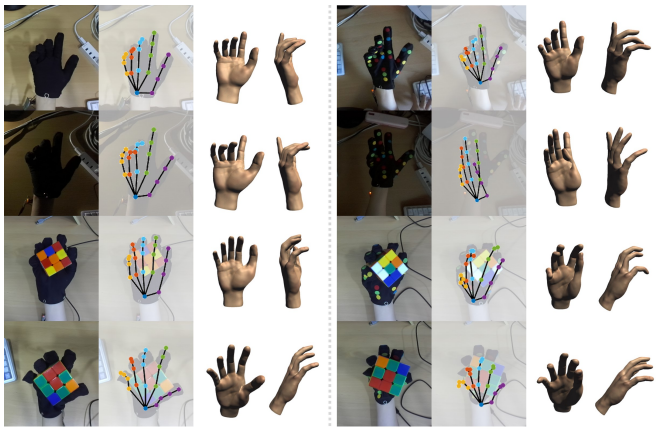

Fig. 8. Tracking results of our framework (left) and VIST (right). The top two figures show hand pose estimations before (1st row) and 5 minutes after (2nd row) the illumination change. The bottom two figures illustrate estimations with object interaction.

marker correspondence is corrupted by the small number of detected markers, leading to inaccurate pose estimation, while our framework directly provides identified keypoints and does not require marker correspondence search. Further evaluations are available in the supplementary video.

C. Ablation study

1) Synthetic pretraining: To study the contribution of pretraining the network with the synthetic dataset, we conduct an ablation study with a quantitative comparison. As Tab. III shows, the pretrained network outperforms the network trained only with the real dataset in terms of accuracy. This can be attributed to the large-scale pretraining that facilitates learning the adequate features to focus on.

TABLE III

QUANTITATIVE EVALUATION FOR ABLATION STUDY (MONOCULAR IMAGES USED). SYNTHETIC PRETRAINING AND IMU-AIDED KEYPOINT INFERENCE ARE ABBREVIATED AS PRETR. AND IMU., RESPECTIVELY.

Pretr.	IMU.	MPKPE	PCK @15mm	PA-MPKPE	PA-PCK @15mm
✗	✗	17.53	0.48	8.82	0.92
✗	✓	19.30	0.41	**7.89**	0.93
✓	✗	17.78	0.49	8.81	0.91
✓	✓	**15.29**	**0.57**	7.96	**0.94**

2) IMU-aided keypoint inference: We also investigate the effect of IMU-aided keypoint inference by comparing the accuracy (Tab. III). Inference without the IMU aid (i.e., inference using the heatmaps output by the network) has larger position errors, largely due to the failure of the network to disambiguate different keypoints in ill-conditioned situations.

VII. CONCLUSION

We present a visual-inertial hand tracking framework where visual information is attained through the lightweight keypoint detection network. We introduce an efficient strategy to generate synthetic and real datasets of gloved hands for training the network. The network predicts heatmaps representing the likelihood of the input image, from which 2D keypoints and uncertainties are estimated via IMU-aided inference using the propagated pose as prior. Tracking is primarily done with IMU signals for high-speed tracking and is refined through factor graph optimization that integrates the predicted keypoint positions, preintegrated IMU signals, and anatomical constraints. The proposed framework is evaluated in challenging scenarios, highlighting its robustness and accuracy. Our framework is directly applicable to a range

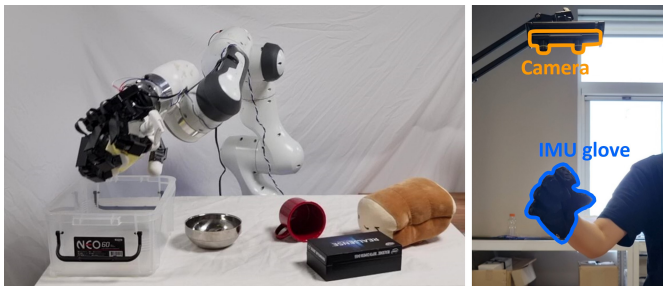

Fig. 9. Application of our hand tracking to robot hand teleoperation.

of domains, including data acquisition for robot learning and teleoperation for dexterous manipulation tasks. The supplementary video demonstrates applications on an interactive simulation using [37] and real-world teleoperation. Future work involves extending the system to bimanual tracking and augmenting the dataset with hand-object interactions to enhance the generalizability of the keypoint detection.

REFERENCES

[1] G. Moon, S. Yu, H. Wen, T. Shiratori, and K. Lee, "InterHand2.6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 548–564.

[2] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges, "Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11 230–11 239.

[3] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2021, pp. 1749–1759.

[4] J. Park, Y. Oh, G. Moon, H. Choi, and K. Lee, "Handoccnet: Occlusion-robust 3d hand mesh estimation network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1496–1505.

[5] H. Xu, T. Wang, X. Tang, and C. Fu, "H2ONet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 17 048–17 058.

[6] Z. Zhou, S. Zhou, Z. Lv, M. Zou, Y. Tang, and J. Liang, "A simple baseline for efficient hand mesh reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1367–1376.

[7] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9826–9836.

[8] S. Sridhar, H. Rhodin, H. Seidel, A. Oulasvirta, and C. Theobalt, "Real-time hand tracking using a sum of anisotropic Gaussians model," in *Proc. Int. Conf. 3D Vis.*, 2014, pp. 319–326.

[9] Y. Li, Z. Xue, Y. Wang, L. Ge, Z. Ren, and J. Rodriguez, "End-to-end 3d hand pose estimation from stereo cameras," in *Proc. Brit. Mach. Vis. Conf.*, 2019.

[10] J. Yang, J. Li, G. Li, H. Wu, Z. Shen, and Z. Fan, "MLPHand: Real time multi-view 3d hand reconstruction via mlp modeling," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 407–424.

[11] S. Han, P. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan *et al.*, "Umetrack: Unified multi-view end-to-end hand tracking for VR," in *Proc. SIGGRAPH Asia*, 2022, pp. 1–9.

[12] C. Wan, T. Probst, L. Gool, and A. Yao, "Self-supervised 3d hand pose estimation through training by fitting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 853–10 862.

[13] J. Malik, S. Shimada, A. Elhayek, S. Ali, C. Theobalt, V. Golyanik, and D. Stricker, "HandVoxNet++: 3d hand shape and pose estimation using voxel-based neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8962–8974, 2021.

[14] P. Ren, Y. Chen, J. Hao, H. Sun, Q. Qi, J. Wang, and J. Liao, "Two heads are better than one: Image-point cloud network for depth-based 3d hand pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 2163–2171.

[15] X. Liu, P. Ren, Y. Gao, J. Wang, H. Sun, Q. Qi, Z. Zhuang, and J. Liao, "Keypoint fusion for rgb-d based 3d hand pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 4, 2024, pp. 3756–3764.

[16] G. Santaera, E. Luberto, A. Serio, M. Gabiccini, and A. Bicchi, "Low-cost, fast and accurate reconstruction of robotic and human postures via imu measurements," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2015, pp. 2728–2735.

[17] T. Baldi, S. Scheggi, L. Meli, M. Mohammadi, and D. Prattichizzo, "GESTO: A glove for enhanced sensing and touching based on inertial and magnetic sensors for hand tracking and cutaneous feedback," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 6, pp. 1066–1076, 2017.

[18] Y. Lee, M. Kim, Y. Lee, J. Kwon, Y. Park, and D. J. Lee, "Wearable finger tracking and cutaneous haptic interface with soft sensors for multi-fingered virtual manipulation," *IEEE/ASME Trans. Mechatron.*, vol. 24, no. 1, pp. 67–77, 2018.

[19] H. Chang and J. Chang, "Sensor glove based on novel inertial sensor fusion control algorithm for 3-D real-time hand gestures measurements," *IEEE Trans. Ind. Electron.*, vol. 67, no. 1, pp. 658–666, 2019.

[20] Y. Liu, C. Lin, and Z. Li, "WR-Hand: Wearable armband can track user's hand," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 3, pp. 1–27, 2021.

[21] G. Park, A. Argyros, J. Lee, and W. Woo, "3d hand tracking in the presence of excessive motion blur," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 5, pp. 1891–1901, 2020.

[22] N. Gosala, F. Wang, Z. Cui, H. Liang, O. Glauser, S. Wu, and O. Sorkine-Hornung, "Self-calibrated multi-sensor wearable for hand tracking and modeling," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 03, pp. 1769–1784, 2023.

[23] Y. Lee, W. Do, H. Yoon, J. Heo, W. Lee, and D. J. Lee, "Visual-inertial hand motion tracking with robustness against occlusion, interference, and contact," *Sci. Robot.*, vol. 6, no. 58, p. eabe1315, 2021.

[24] X. Wang, P. Ren, H. Zhang, X. Sheng, D. Li, L. Xie, Y. Gao, and E. Yin, "Vihand: Enhancing 3d hand pose estimation with visual-inertial benchmark," in *Proc. ACM Int. Conf. on Multimedia*, 2025, pp. 12 753–12 760.

[25] M. Trumble, A. Gilbert, C. Malleson, A. Hilton, and J. Collomosse, "Total Capture: 3D human pose estimation fusing video and inertial sensors," in *Proc. Brit. Mach. Vis. Conf.*, vol. 2, no. 5, 2017, pp. 1–13.

[26] Z. Zhang, C. Wang, W. Qin, and W. Zeng, "Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2200–2209.

[27] H. Liang, Y. He, C. Zhao, M. Li, J. Wang, J. Yu, and L. Xu, "Hybridcap: Inertia-aid monocular capture of challenging human motions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 2, 2023, pp. 1539–1548.

[28] Epic Games, "Unreal Engine," https://www.unrealengine.com, 2023.

[29] E. Games, *Chaos Physics*, 2024, available at https://dev.epicgames.com/documentation/en-us/unreal-engine/physics-in-unreal-engine.

[30] P. Denis, J. Elder, and F. Estrada, "Efficient edge-based methods for estimating Manhattan frames in urban imagery," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 197–210.

[31] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. of Comput. Vis.*, 2022, pp. 2149–2159.

[32] S. Chatterjee, D. Doan, and B. Calli, "Utilizing inpainting for training keypoint detection algorithms towards markerless visual servoing," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 3086–3092.

[33] P. Vasu, J. Gabriel, J. Zhu, O. Tuzel, and A. Ranjan, "FastViT: A fast hybrid vision transformer using structural reparameterization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 5785–5795.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5693–5703.

[35] N. Trawny and S. Roumeliotis, "Indirect kalman filter for 3D attitude estimation," Univ. Minnesota, Dept. Comput. Sci. Eng., Tech. Rep., 2005.

[36] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation," in *Proc. Robot.: Sci. Syst.*, 2015.

[37] H. Ji, H. Kim, J. Lee, S. Lee, S. An, J. Heo, Y. Lee, Y. Lee, and D. J. Lee, "GPU-accelerated subsystem-based ADMM for large-scale interactive simulation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2025.