Multi-Stage Ensemble Optimization for Articulated 2D-3D Medical Image Registration in Image-Guided Wrist Surgery

Ata Soloukibashiz[a], Gianni Allebosch[a], Matthias Vanhees[b], Hiep Quang Luong[a], Peter Veelaert[a], and Brian G. Booth[a]

[a]TELIN-IPI, Ghent University – imec
[b]Department of Orthopedic Surgery, AZ Monica Hospital, Monica Orthopedic Research (MoRe) Institute, and Antwerp University Hospital, Antwerp, Belgium

ABSTRACT

The registration of articulated 3D models to 2D images is a fundamental challenge in image-based guidance of joint surgery. In this context, optimization-based approaches have long been popular as they are immediately applicable to new scenarios without requiring extensive pre-training using large datasets. However, optimization-based registration methods are often hindered by the risk of converging to local minima, particularly when dealing with articulated structures, partial occlusions, or objects with symmetries.

To address the local minima issue, we present a novel multi-stage ensemble optimization strategy designed specifically to be more resilient to local minima. Inspired by block coordinate descent, our approach consists of multiple stages in which, at each stage, an ensemble of optimizers operates in parallel on subsets of pose parameters. These subsets form lower-dimensional parameter domains, which represent the range of possible values for the pose parameters (e.g., translation and rotation). By focusing on smaller, simplified subspaces, the optimization landscape becomes smoother, reducing abrupt changes in the loss function and improving convergence. The combination of multiple optimization trajectories and simplified parameter domains increases the likelihood of finding the global optimum and achieving faster convergence compared to methods that optimize all parameters simultaneously.

We evaluated our method by registering a preoperative 3D CT wrist model with intra-operative optical images acquired during image-guided Percutaneous Scaphoid Fixation (PSF) surgery. The registration accuracy of our technique exceeded competing optimization techniques, specifically our optimized pose parameters achieved root mean squared errors between 13.59% and 33.95% lower than those of competing methods. These results suggest the broader applicability of our optimization strategy for articulated registration in various surgical 2D-3D alignment problems.

Keywords: Wrist pose estimation, Surgical guidance, Articulated registration, Ensemble Optimization, Multi-Stage Optimization

1. INTRODUCTION

The registration of a 3D model to 2D images is a fundamental challenge in various domains, including image-guided surgery, human body pose estimation, and object registration. In the case of wrist surgery, the alignment involves transforming a pre-operative 3D wrist model, which is reconstructed from pre-surgical imaging modalities such as CT scans, so that its rendered 2D projections match intra-operative 2D images acquired during the procedure. This process, as illustrated in Figure 1, inherently involves projecting the 3D model into a lower-dimensional 2D space, which inevitably discards some spatial information. Consequently, when registering the projected model through direct optimization, the cost function guiding the optimization lacks the complete geometric context of the original 3D structure, leading to a more fragmented optimization landscape with potential local minima. Furthermore, the articulation of joints, self-occlusions, and object symmetries compound these challenges, making it even harder to reliably locate the global optimum. These difficulties are particularly significant in wrist surgery, where precise registration is essential for accurate surgical navigation. Poor 2D-3D registration can adversely impact surgical outcomes, highlighting the importance of robust optimization techniques tailored to this problem.

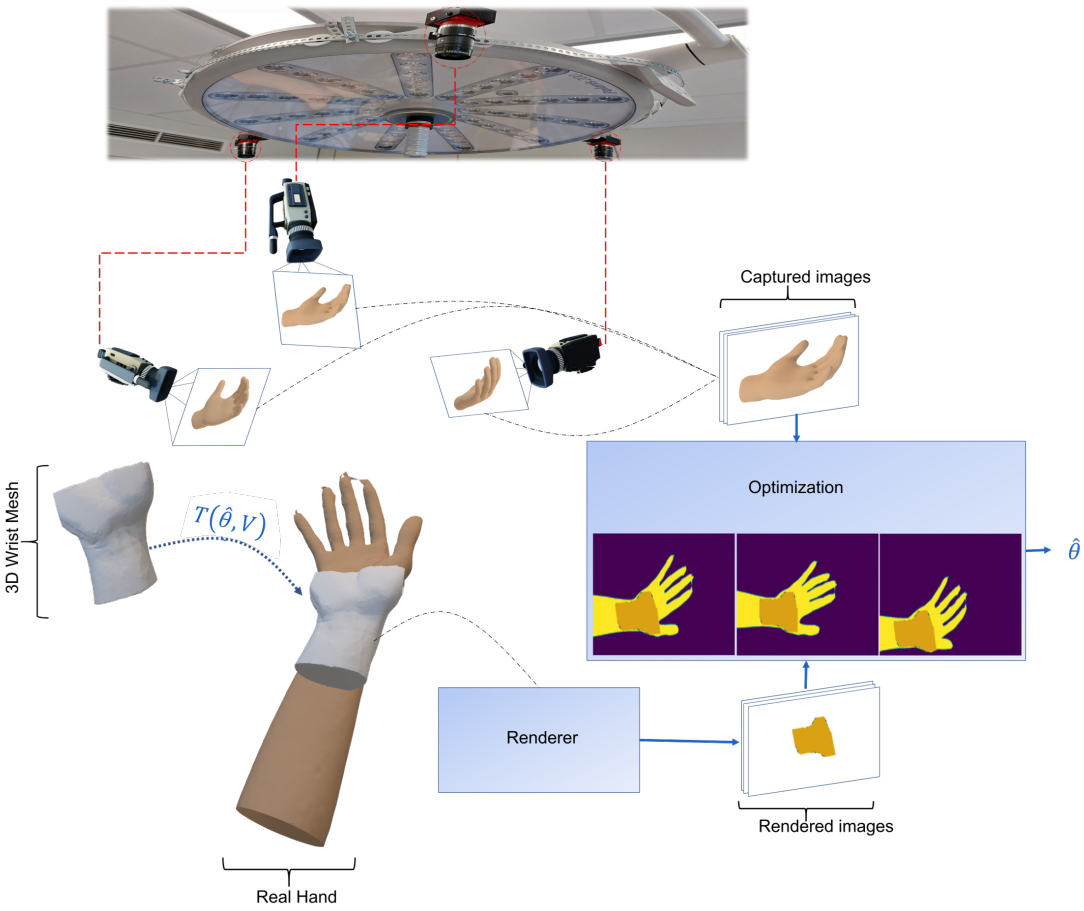


Figure 1: Overview of the proposed wrist pose estimation system. Multiple near-infrared (NIR) cameras capture images of the patient's wrist during surgery. The acquired images are segmented, then used to align a pre-operative 3D wrist model to the patient's wrist. This alignment is performed by aligning the rendered silhouette of the 3D model with the segmented wrist in each image.

With the high-accuracy demands of image-guided surgery in mind, 2D-3D registration techniques have been widely studied. Several works proposed both optimization-based and neural network-based approaches to align 2D image data with articulated 3D models. The approaches of Bogo et al.[1] and Wu et al.[2] exemplify optimization-based techniques, while others, such as Dwivedi et al.,[3] Kanazawa et al.[4] and Khaleghi et al.[5] leverage deep learning models trained on large datasets to achieve effective registration. This persistent division in the literature between direct optimization methods and deep learning approaches highlights the challenges involved in collecting a sufficiently large training dataset for the extensive deep learning training task.[6] As a result, direct optimization methods are still commonly used for registration tasks in image-guided wrist surgery.[7]

For the direct optimization of 2D-3D registration problems, the non-linearity and non-convexity of the optimization landscape present a major challenge. To navigate this challenge, researchers have proposed various ways to smooth out the optimization landscape. One such technique is differentiable rendering, which facilitates efficient gradient-based optimization by enabling a smoother optimization landscape. Techniques such as the soft rasterizer (SoftRas), used by Liu et al.[8] and Dwivedi et al.,[3] provide differentiable rendering to maintain efficient gradient flow even in complex articulated poses. In addition, Wu et al.[2] demonstrated an analytical differentiable renderer, enhancing efficiency compared to numerical gradient methods. These differentiable rendering approaches have proven to be particularly effective in offering robust alignment capabilities.

To further address the optimization challenges, several studies have utilized multi-stage optimization techniques, techniques that divide larger optimization problems into sequence of simpler optimization "stages", with

each stage usually optimizing a subset of the pose parameters. An advantage of multi-stage optimization is that it allows for a systematic approach to pose parameter refinement. Bogo et al.[1] introduced a multi-stage strategy for whole body pose alignment. They began with the optimization of global pose parameters, such as torso alignment, before refining joint-specific pose parameters, effectively reducing the risk of local minima. This concept of multi-stage refinement was also adopted by Song et al.,[9] who combined neural network predictions with gradient-based refinement, thus enhancing robustness while maintaining precision. Balan and Black[10] employed a similar strategy to balance computational efficiency with the accuracy desired for joint-specific optimizations.

Ensemble methods have also been effective in tackling the challenges inherent in the registration of complex articulated structures. Tremblay et al.[11] adopted an ensemble approach, using multiple parallel optimizations with varying random initializations and learning rates to prevent convergence towards a local minima. This ensemble strategy, particularly beneficial for symmetric structures, increased the probability of finding a global solution without exhaustive parameter tuning, demonstrating a promising direction for challenging registration problems.

While differentiable rendering, multi-stage optimization, and ensemble optimization have shown to be valuable techniques in 2D-3D image registration tasks, their benefits have yet to be combined, or applied to the registration challenges involved in image-guided wrist surgery. In our proposed method, we integrate insights from multi-stage optimization, ensemble techniques, and differentiable rendering to register an articulated 3D wrist model to intra-operative 2D images (see Fig. 1). Our method aims to reduce convergence issues by systematically refining subsets of wrist pose parameters, thereby improving robustness and efficiency. The ultimate goal is to enhance intra-operative visualization, providing surgeons with a precise, real-time augmented representation of the internal wrist anatomy, even in challenging conditions involving self-occlusion and articulation. By combining these advanced optimization strategies, our approach strives to address the limitations of existing methods and offer a solution with improved accuracy and reliability for surgical applications. The final aim is to achieve sub-millimeter accuracy for accurate visualization of internal bone structures within an image-guided surgical environment.

2. MATERIALS

In this study, a pre-operative 3D wrist model was made from a CT scan of the right wrist of a healthy volunteer. The scan was acquired with a resolution of $0.25 \times 0.25 \times 0.25$ mm and a field of view of $128 \times 128 \times 90.75$ mm.

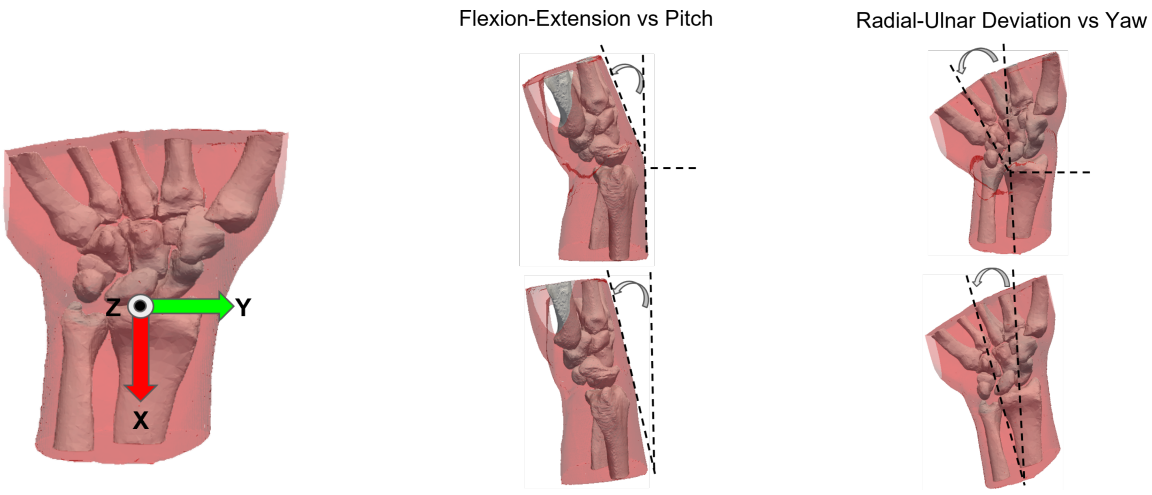
For intra-operative imaging, we utilized a multi-camera system consisting of three monochrome Allied Vision Alvium 1800 U-501 NIR cameras, equipped with 800 nm low-pass filters to limit image overexposure due to the brightness of the surgical light (see Fig. 1). These cameras captured high-resolution images at 25 fps with a resolution of 1920×1080 pixels, covering a field of view of 53×28 cm. A hardware trigger was used to synchronize the image capture across all three cameras. The exposure time of each camera was empirically set to 5000 µs.

3. METHOD

To achieve an accurate 2D-3D registration of the wrist, we propose a method that combines an anatomically accurate 3D wrist model with a robust optimization framework. An outline of the registration method is shown in Fig. 1. The method aims to match 2D renderings of the 3D wrist model to 2D segmentations of the wrist captured by the intra-operative cameras. The subsections below outline the key components of our approach in detail.

3.1 3D Wrist Model

The 3D wrist model was constructed using a CT scan which were manually segmented to create an anatomically accurate mesh representation of the wrist. The segmentation process involved identifying and extracting critical anatomical regions, including the skin surface, the radius, ulna, and all carpal and metacarpal bones. The skin surface and each bone were saved as 3D triangular surface meshes. Linear Blend Skinning (LBS)[12] was then applied to the skin surface mesh, thereby link the motion of the skin surface to the motion of the individual bones meshes. In this approach, the bones were treated as rigid bodies, while vertices on the skin mesh are transformed



(a) (b) (c)

Figure 2: (a) 3D wrist model with the coordinate system centered at the wrist joint. (b) Illustration of different y-axis rotations: the wrist flexion/extension (top) and the arm's pitch rotation (bottom). (c) Illustration of different z-axis rotations: the wrist radial/ulnar deviation (top) and the arm's yaw rotation (bottom).

using a linear combination of rigid motions from nearby bones. For the purposes of registration, only the skin surface is utilized as it directly interacts with the 2D image data, while the bones and other anatomical structures serve as supplementary components to enhance visualization. We will denote this 3D wrist model as:

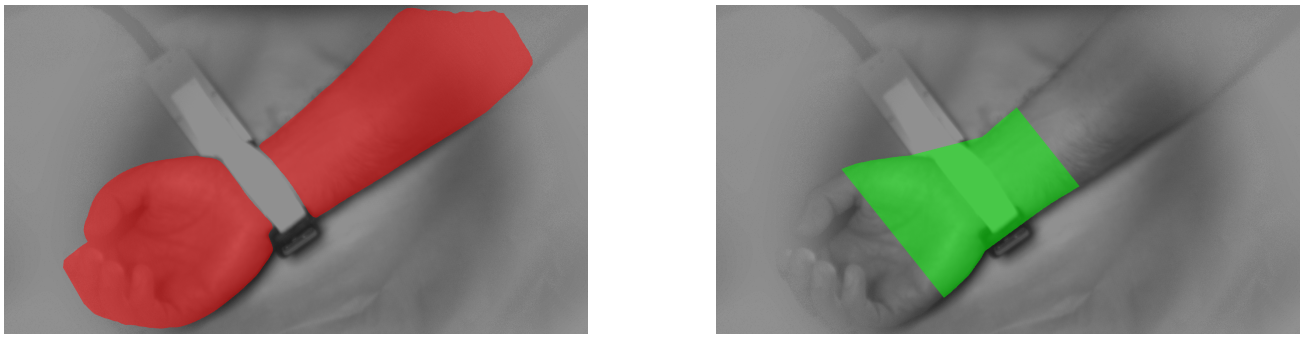
$$M = (F, \mathbf{V}) \tag{1}$$

where $\mathbf{V}$ is a matrix of vertex coordinates in 3D space, and F is an array of triangular faces, each defined by indices referencing $\mathbf{V}$. Together, M encapsulates both the geometric structure (via $\mathbf{V}$) and the topological relationships (via F) of the 3D triangular mesh, serving as a compact representation of the model

To define the wrist's pose parameters, a coordinate system was established at the wrist joint, as recommended by Wu et al.[13] (Fig. 2(a)). While the translations are computed in a global (fixed) coordinate system, the rotations are defined in a body-fixed (local) coordinate system to ensure that each rotation is applied sequentially with respect to the current orientation of the coordinate system itself (see Fig. 2(a)). This enabled the wrist meshes to model motions from its two primary degrees of freedom: flexion-extension and radial-ulnar deviation (see Fig. 2(b, c)). Both movements were implemented with motion constraints derived from the anatomical structure to reflect the biological limits of the wrist joint, ensuring both realistic and anatomically correct behavior during articulation.[13,14] In total, the 3D wrist model has eight pose parameters:

$$\boldsymbol{\theta} = [t_x, t_y, t_z, \alpha, \beta, \gamma, \phi_{\text{FE}}, \phi_{\text{RU}}], \tag{2}$$

where $t_x, t_y, t_z \in \mathbb{R}^3$ are the global translation parameters, constrained as $t_x, t_y \in [-1\ m, 1\ m]$ and $t_z \in [0.5\ m, 1.5\ m]$. These ranges reflect the practical limits of wrist movement during surgical procedures and ensure that the wrist remains within the field of view (FOV) of the camera system. α, β, γ are Euler angles representing the wrist's orientation ($\alpha \in [-180°, 180°]$ (roll), $\beta \in [-90°, 90°]$ (pitch), and $\gamma \in [-180°, 180°]$ (yaw)). $\phi_{\text{FE}} \in [-70°, 75°]$ is the wrist's flexion-extension angle, and $\phi_{\text{RU}} \in [-20°, 35°]$ is the wrist's radial-ulnar deviation angle. The range of β is constrained to $[-90°, 90°]$ as it reflects the practical limits of wrist motion during surgical procedures, while the ranges for ϕ_{FE} and ϕ_{RU} are anatomically and biomechanically determined to ensure realistic modeling of wrist articulation. It is these eight pose parameters that will be optimized in our registration.



(a) (b)

Figure 3: (a) Initial segmentation of the arm and hand using YOLOv9. The region marked in red highlights the incomplete segmentation, where gaps appear due to occlusions caused by the ultrasound probe. (b) Refined segmentation of the wrist region using a 3D wrist model to crop the segmentation and fill these gaps, ensuring a more complete representation of the wrist.

3.2 Image Enhancement and Segmentation

The intra-operative images collected in our study showed relatively low image quality due to the strength and flickering of the LED surgical light. These lighting artifacts degrade video quality and hinder our downstream task of pose detection. To address these issues, the real-time video enhancement algorithm developed by Allebosch et al.[15] was used to remove these flicker and spotlight artifacts.

Following the enhancement, the wrist region was segmented using a three-stage approach to generate precise binary masks. Initially, a pre-trained YOLOv9[16] instance segmentation network was used to provide preliminary segmentations for the hand and arm region. These segmentations are then refined by filling in gaps and cropping the segmentations to the wrist region modeled in the 3D wrist model. These refinements were performed with the help of geometric constraints such as prior camera calibration data, the known positioning of the arm on the surgical table, the principal axis of the arm determined from PCA of the initial segmentation, and the size constraints of the 3D wrist model.

To further enhance the segmentation quality, temporal smoothing was applied to reduce noise and to ensure frame-to-frame consistency. From a previous frame's segmentation, morphological operations are used to generate minimal and maximal masks. These masks are combined with the refined mask from the current frame to produce the final segmentation output. Specifically, any pixel in the current mask that lies outside the maximal mask is excluded from the segmented hand. Conversely, any pixel within the minimal mask that is not already in the current mask is incorporated into it. This smoothing step is applied to all frames except the first, ensuring accurate and consistent segmentation results for registration tasks (see Fig. 3).

3.3 2D-3D Registration

3.3.1 Objective Function & Optimization

The objective of our 2D-3D registration method is to find the optimal 3D pose parameters $\boldsymbol{\theta}^*$, that align our 3D wrist model (Fig. 2) with the patient's actual wrist as seen in the 2D intra-operative images. To achieve this objective, we employ a gradient-based optimization approach, inspired by render-and-compare techniques,[11,17,18] to refine the pose of the 3D wrist model across multiple stages. In this framework, the alignment is driven by an objective function $L(\boldsymbol{\theta}, \mathbf{V})$, which evaluates the dissimilarity between the rendered projections of the transformed 3D wrist model and the actual silhouette images of the patient's wrist. This objective function is defined as:

$$L(\boldsymbol{\theta}, \mathbf{V}) = \sum_{i=1}^{3} D(\rho_i(\mathbf{F}, T(\boldsymbol{\theta}, \mathbf{V})), \mathbf{I}_i), \tag{3}$$

where the function $T(\boldsymbol{\theta}, \mathbf{V})$ performs a non-rigid transformation that applies the pose parameters, $\boldsymbol{\theta}$, to the vertices, $\mathbf{V}$, of the 3D wrist model, the rendering function, ρ_i, generates a binary silhouette image of the

transformed 3D wrist model as viewed from the i^{th} camera, and $\mathbf{I}_i$ denotes the binary segmented image from the same camera. The dissimilarity function D compares the rendered image with the segmented image. In our approach, we use the Jaccard Index[19] as the dissimilarity measure, and a differentiable renderer for ρ_i implemented using a soft rasterizer.[8] The differentiable rendering enables gradient propagation through the rendering pipeline, allowing the gradients of the loss function $L(\boldsymbol{\theta}, \mathbf{V})$ to be analytically computed and used by gradient descent optimizers.

The estimated wrist pose, $\hat{\boldsymbol{\theta}}$, is defined as one that minimizes the objective function L in Equation 3:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ L(\boldsymbol{\theta}, \mathbf{V}) \quad \text{subject to} \quad \mathbf{lb} \leq \boldsymbol{\theta} \leq \mathbf{ub}, \tag{4}$$

where $\mathbf{lb} = [-1, -1, 0.5, -180, -90, -180, -70, -20]$ and $\mathbf{ub} = [1, 1, 1.5, 180, 90, 180, 75, 35]$ are vectors containing the previously-defined lower and upper bounds of the pose parameters $\boldsymbol{\theta}$.

3.3.2 Multi-Stage Optimization

Directly solving this objective function in a single step is challenging due to the non-linearity introduced by the 3D-to-2D projection, which can potentially lead to local minima.[4] To mitigate this challenge, a multi-stage optimization approach, inspired by block coordinate descent (BCD),[20] is adopted. In this multi-stage approach, the optimization problem is divided into N smaller sub-problems that sequentially optimize subsets of parameters, $\boldsymbol{\psi}$. We refer to these parameter subsets as blocks and we define our optimal 3D wrist pose parameters as the sum of these parameter blocks across all optimization stages:

$$\hat{\boldsymbol{\theta}} = \sum_{j=1}^{N} \hat{\boldsymbol{\psi}}_j. \tag{5}$$

where each block, $\boldsymbol{\psi}$, is a vector of the same size as $\boldsymbol{\theta}$ but with the non-optimized parameters fixed to zero. For each stage j, the optimization of the block parameters follows the same objective function minimization as defined earlier:

$$\hat{\boldsymbol{\psi}}_{\mathbf{j}} = \underset{\boldsymbol{\psi}}{\operatorname{argmin}}\ L(\boldsymbol{\psi}, \mathbf{V}_{\mathbf{j-1}}) \quad \text{subject to} \quad \mathbf{lb}_j \leq \boldsymbol{\psi} \leq \mathbf{ub}_j, \tag{6}$$

where $\mathbf{lb}_j[k] = \mathbf{ub}_j[k] = 0$ for indices k corresponding to pose parameters in $\boldsymbol{\theta}$ that are not optimized in stage j, and the remaining lower and upper bounds are adjusted at each optimization stage based on the estimated pose parameters of previous stages. The vertices of the 3D wrist model, $\mathbf{V}_{j-i}$, are also transformed using the pose parameters estimated from previous stages.

3.3.3 Multi-Stage Ensemble Optimization (MSEO)

Finally, the choice of parameter blocks remains to be defined. While multi-stage optimization methods have been used in medical image registration (e.g., rigid → affine → deformable), they often rely on manual grouping of pose parameters based on domain knowledge. Such manual groupings may be sub-optimal because the interactions between pose parameters can influence the loss function in unpredictable and subject-specific ways. To overcome these limitations, we propose using a parallel ensemble of gradient descent optimizers at each stage.[11] Each optimizer estimates values for a different, randomly-selected subset of pose parameters. Once all optimizers have converged, the pose parameters that produce the minimum loss across the ensemble are retained and used to update the optimization process. The full multi-stage ensemble optimization (MSEO) process is presented in Algorithm 1.

Each optimizer in the ensemble uses an Adam optimizer with an initial learning rate of 0.002. The learning rate is adapted dynamically during optimization by monitoring the loss function. If the loss remains unchanged for 4 consecutive iterations, the learning rate is reduced by a factor of 0.2. This process continues until the learning rate falls below the termination threshold of 0.00001. Convergence is determined to have occurred when the learning rate falls below this threshold. Our MSEO method was implemented using the PyTorch3D framework[21] on a high-performance workstation equipped with an NVIDIA GeForce RTX 3080 Ti GPU and using CUDA to accelerate the rendering and optimization processes.

Algorithm 1 Multi-Stage Ensemble Optimization for Articulated 2D-3D Image Registration

Input: 3D wrist model faces F and vertices $\mathbf{V}_0$, initial pose parameters $\boldsymbol{\theta}_0$, segmented wrist images $\mathbf{I}_i$ (from cameras $i = \{1, 2, 3\}$)
Output: Estimated pose parameters $\hat{\boldsymbol{\theta}}$
for each stage $j = 1, \cdots, N$ **do**
  Update $\mathbf{lb}_j$ and $\mathbf{ub}_j$ based on $\boldsymbol{\theta}_{j-1}$
  for each optimizer p in ensemble **do**
    Generate random binary 8-vector $\mathbf{m}_p$
    $\mathbf{lb}_{j,p} \leftarrow \mathbf{lb}_j \odot \mathbf{m}_p$
    $\mathbf{ub}_{j,p} \leftarrow \mathbf{ub}_j \odot \mathbf{m}_p$
    $\hat{\boldsymbol{\psi}}_{j,p} \leftarrow \underset{\boldsymbol{\psi}}{\operatorname{argmin}}\, L(\boldsymbol{\psi}, \mathbf{V}_{j-1})$ subject to $\mathbf{lb}_{j,p} \leq \boldsymbol{\psi} \leq \mathbf{ub}_{j,p}$, (Eq. 6)
    Calculate loss $\ell_{j,p} \leftarrow L(\hat{\boldsymbol{\psi}}_{j,p}, \mathbf{V}_{j-1})$, (Eq. 1)
  end for
  Select block $\hat{\boldsymbol{\psi}}_j$ with minimum loss $\ell_j^* = \min_p \ell_{j,p}$
  $\boldsymbol{\theta}_j \leftarrow \boldsymbol{\theta}_{j-1} + \hat{\boldsymbol{\psi}}_j$, (Eq. 5)
  $\mathbf{V}_j \leftarrow T(\boldsymbol{\theta}_j, \mathbf{V}_0)$
end for
Return $\boldsymbol{\theta}_N$

4. EXPERIMENTS

To evaluate the performance of our MSEO registration method, we conducted experiments on both synthetic and real datasets, each serving a distinct purpose. These datasets were chosen to comprehensively assess the registration accuracy, robustness, and computational efficiency of MSEO under both controlled and realistic conditions.

For both the synthetic and real datasets, comparisons were made against four established optimization strategies to benchmark the MSEO's performance:

1. **Simultaneous Optimization (SO):**[4] All pose parameters were optimized simultaneously using the Adam optimizer.
2. **Rigid-Articulated Grouping (RAG):**[10] Parameters were grouped into rigid (translation and rotation) and articulated (joint angles) blocks, with a final stage optimizing all parameters together.
3. **Intuitive Grouping (IG):**[1] Parameters were grouped using domain knowledge and optimized in the following order: (a) translation, (b) yaw and radial/ulnar deviation, (c) pitch and flexion/extension, (d) roll, and (e) all parameters together.
4. **Diff-Dope**:[11] Optimizes all parameters concurrently with Adam using varying learning rates, with the results being aggregated for enhanced robustness.

4.1 Synthetic Data Experiment

To evaluate MSEO's registration accuracy with known ground truth pose parameters, a synthetic data set of 36 random wrist poses was generated with the help of the 3D wrist model. Each scenario in this dataset was created by placing the 3D wrist model in a random pose and rendering this pose into each of the three cameras. These renderings are then used as the segmentations $\mathbf{I}_1, \mathbf{I}_2, \mathbf{I}_3$. Subsequently, a random transformation, $\boldsymbol{\theta}$, was applied to the 3D wrist model: translations were applied to the wrist model in the t_x, t_y, and t_z axes, with displacements ranging approximately between -100mm and 100mm units; rotational perturbations were introduced using Euler angles (α, β, γ), with angular variations between $-11.5°$ and $11.5°$; and the wrist joint angles, representing flexion-extension (ϕ_{FE}) and radial-ulnar deviation (ϕ_{RU}), were adjusted within the range of

$-8.6°$ to $11.5°$. These transformations ensured the synthetic dataset captured plausible wrist movements while maintaining anatomical realism.

The goal of the synthetic data experiment is then to align the transformed 3D wrist model to its original renderings. The resulting pose parameters are then compared to the ground truth transformation, which in this case is $-\boldsymbol{\theta}$, using the root mean squared error (RMSE) metric.

4.2 Real Data Experiment

We extended our evaluation using a real dataset comprising of 160 image triplets captured by the multi-camera system described in the Materials section. The data was collected from eight mock percutaneous scaphoid fixation surgeries in a surgical environment at the operating room of AZ Monica Hospital (Antwerp Campus). A total a 4000 image triplets were recorded, but most images showed similar poses due to limited patient motion. To balance computational efficiency with sufficient temporal variability, one image triplet out of every 25 was sampled from the complete sequence (i.e., 160 of the 4000 recorded image triplets), thereby reducing redundancy in the dataset as well as the processing load. To initialize the registration algorithm, we take advantage of the fact that the 160 image triplets were recorded sequentially, a manual initialization was performed for the first image triplet, while subsequent image triplets were initialized using the registration result from the previous image triplet.

For this real dataset, where ground truth annotations were unavailable, we assessed the performance of our method (MSEO) using four complementary loss metrics: Jaccard Loss (i.e., the loss function used in all optimizations), Chamfer Loss,[22] Dice Loss,[23] and Pixel Correspondence Metric (PCM)[24] Loss. These metrics were carefully selected to capture diverse aspects of performance, including pixel-level similarity,[25] edge-based similarity,[22] and overlap-based similarity.[23] Statistical t-tests are then used to compare the loss function values of MSEO against the competing optimization methods.

5. RESULTS

5.1 Synthetic Dataset Results

The estimated pose parameters for the 36 synthetic wrist poses were compared to the ground truth values using the root mean squared error (RMSE) metric, which quantified the alignment accuracy. The results, summarized in Table 1, show that our method, MSEO, generally outperforms competing optimization strategies, achieving improvements ranging from 13.59% to 33.95%. The lone exception is with the flexion-extension angle, where Rigid-Articulated Grouping (RAG) performed slightly better. On average MESO provided registrations with less than 3 mm of translation error and $10°$ of rotational error.

Table 1: RMSE values for different methods on a synthetic dataset of 36 random wrist poses. For translations, values are in millimeters (mm), and for rotations, values are in degrees (°).

Parameter	**SO**	**RAG**	**IG**	**Diff-Dope**	**MSEO**
Translation (x) [mm]	9.96	8.60	10.58	2.93	**2.87**
Translation (y) [mm]	3.46	3.41	3.36	3.35	**2.63**
Translation (z) [mm]	18.68	13.78	17.38	3.67	**2.59**
Rotation (roll) [°]	6.28	6.48	5.95	6.76	**5.81**
Rotation (pitch) [°]	6.33	6.04	5.73	5.90	**4.90**
Rotation (yaw) [°]	10.36	9.76	9.02	9.89	**6.73**
Flexion / extension [°]	9.10	**8.78**	8.89	9.36	8.91
Radial / longitudinal deviation [°]	8.73	8.32	8.12	7.90	**7.67**

5.2 Real Dataset Results

Fig. 4 shows the distributions for all four loss functions on the 160 image triplets on the real dataset. For all four loss functions, MSEO achieves a significantly lower loss than competing methods. Subsequent t-tests confirmed that these improvements by MSEO are statistically significant across all loss functions, with the highest observed p-value being 5.3×10^{-36}, and the remaining p-values even lower.

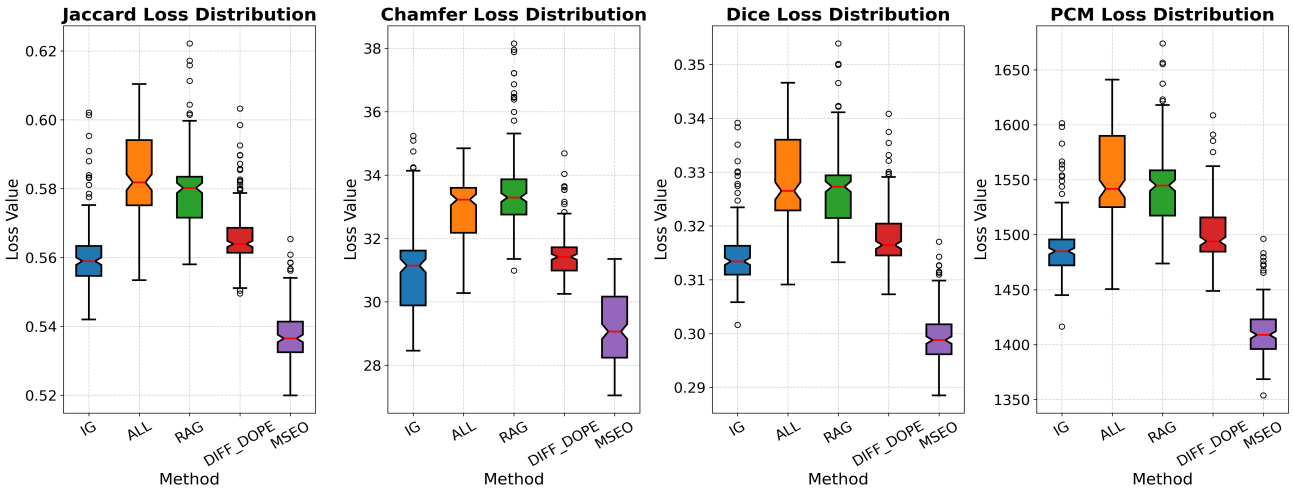


Figure 4: Boxplot comparison of the loss values obtained using different optimization methods across the 160 image triplets in the real image dataset. Each box represents the interquartile range (IQR), spanning from the 25th percentile (lower boundary) to the 75th percentile (upper boundary), with the line inside indicating the median (50th percentile). The whiskers extend to the most extreme data points within 1.5 times the IQR, while the circles represent outliers beyond this range. Note that MSEO significantly outperforms competing optimizers on the dataset.

In Fig. 5 we show the loss function results across all 160 image triplets in our dataset. We observed that MSEO achieved superior performance in 100% of image triplets for the Jaccard, Dice, and PCM Loss metrics, while achieving the lowest loss values in 88.96% of image triplets for the Chamfer Loss. These results show a registration improvement by MSEO that is generally consistent across wrist poses.

Finally, Fig. 6 presents a qualitative comparison of alignment results before and after applying MSEO. The alignment improvements illustrate MSEO's effectiveness in adapting to real-world variations and achieving a good wrist alignment, even in the absence of ground truth annotations.

6. DISCUSSION

Our results demonstrate that the proposed method outperforms existing approaches in aligning a 3D wrist model with patient anatomy, achieving higher accuracy in registration. However, as outlined in the introduction, the ultimate goal is to achieve real-time, sub-millimeter accuracy to enable precise visualization of internal bone structures in image-guided surgical environments. While our method marks significant progress, it does not yet meet this stringent requirement, highlighting the need for further advancements.

Nevertheless, one of the primary strengths of our method lies in its direct optimization framework, which eliminates the need for annotated AI model training datasets. This feature makes the method robust in scenarios where obtaining high-quality labeled data is challenging or infeasible. The ability to directly optimize in a 2D-3D space also enables flexibility and opens up avenues for future research. For instance, this approach allows for exploring smarter grouping and sequencing of parameters during optimization, potentially improving both efficiency and accuracy. We also obsrved that our MSEO registration was robust to certain types of noise, such as segmentation inaccuracies, thanks to its reliance on silhouette-based alignment rather than precise point correspondences. Furthermore, the differentiable rendering framework combined with ensemble and multi-stage optimization introduces a modular design that can be expanded or customized to specific applications, making it a versatile tool for 3D-2D registration tasks.

Despite its strengths, the method has several limitations that can be addressed in future work. The first, the linear blend skinning used in the 3D wrist model is only an approximation of the real geometrical changes that occur on the skin surface due to wrist motion. This approximation can introduce inaccuracies in the registration process. Additionally, the method's performance, like those of competing methods, is sensitive to initialization:

poor initial estimates of the pose parameters can lead to suboptimal alignment or even failure to converge. Moreover, the silhouettes generated by the soft rasterizer, and those obtained through segmentation, are not perfectly matched even when ground-truth data is used. This discrepancy is due to differences in quantization and discretization as the soft-rasterizer provides a smoothed wrist silhouette whereas the image segmentation produces a crisp silhouette. These discrepancies can adversely affect the optimization process. So too can segmentation errors in the image processing. Accurate segmentation of the wrist region also remains a challenge, as accurately distinguishing the wrist region from the surrounding hand anatomy is not trivial.

However, the most notable limitation may be the algorithm's computation time, which currently runs at approximately 20 minutes per registration. Rendering, a core component of the method, is computationally expensive. The multi-stage, ensemble-based approach we employ involves iterative rendering across several stages, which significantly increases computational time. For real-time applications, such as providing surgeons with updated bone poses during surgery, this latency is a critical bottleneck.

While our method has limitations, there are clear pathways for improvement. For example, integrating auxiliary data sources, such as real-time ultrasound tracking,[26] can help refine bone pose estimation by capturing subtle movements that are difficult to model with a purely silhouette-based approach. Additionally, improving the skinning in the 3D wrist model to better represent the complex changes in wrist geometry due to motion could also produce improvements. In this respect, the statistical modeling of wrist geometry may prove useful.[27,28] Similarly, incorporating advanced initialization strategies could mitigate some of the current shortcomings for this and other 2D-3D image registration techniques.

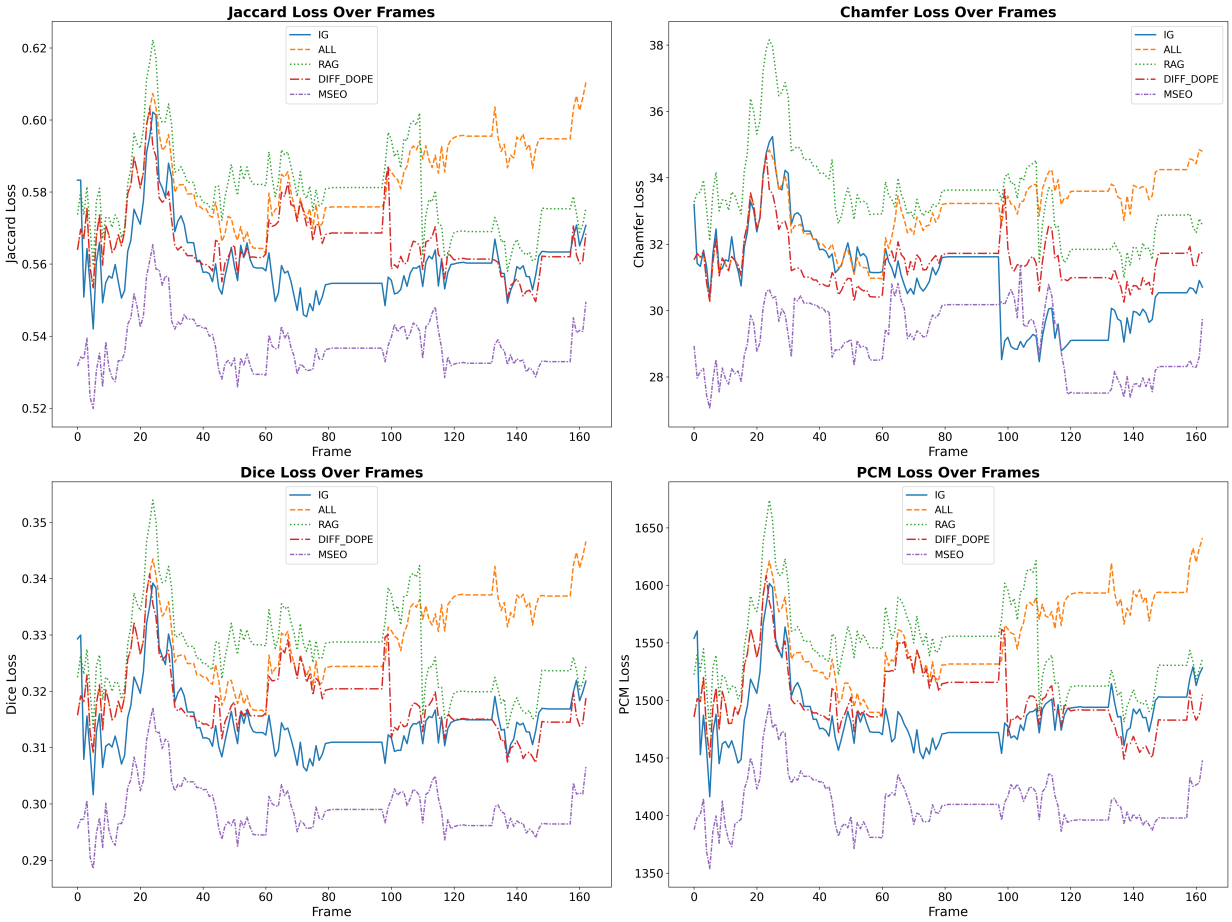


Figure 5: Plot of the loss values of each optimization method for each of the 160 image triplets in the real image dataset. Note that MSEO consistently outperforms competing optimizers on this dataset.

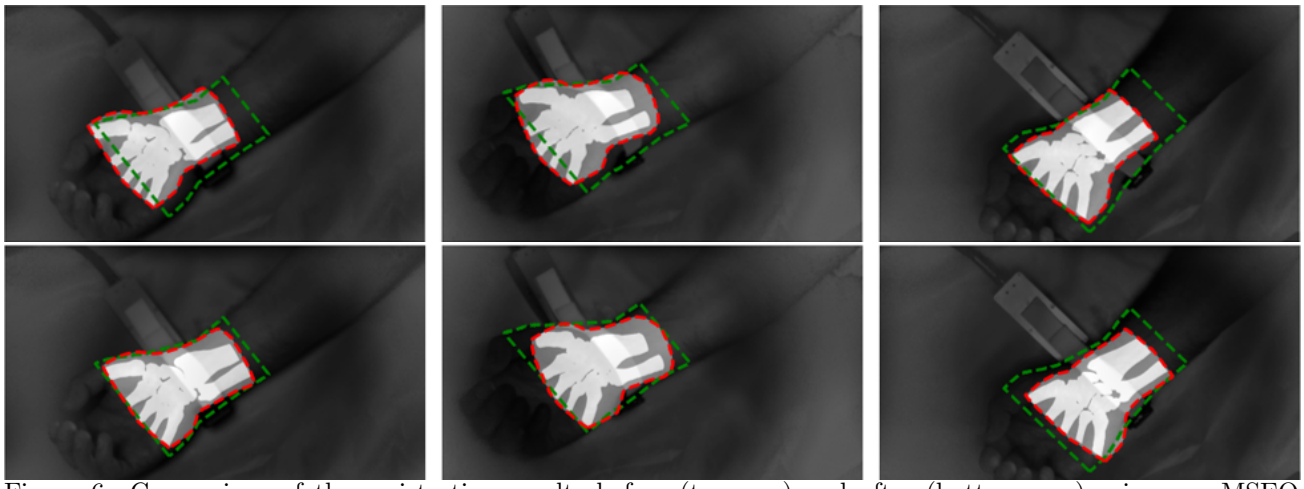


Figure 6: Comparison of the registration results before (top row) and after (bottom row) using our MSEO method. The green dashed line represents the segmentation contour, and the red dashed line represents the rendered mesh contour. Note that MSEO provides a reasonable result even in the presence of segmentation errors.

Future work could also explore alternative optimization strategies to further reduce sensitivity to local minima, such as hybrid approaches combining gradient descent with global optimization techniques, like genetic algorithms. Finally, optimizing the efficiency of the rendering process, either through algorithmic improvements or hardware acceleration, could make this method more suitable for real-time surgical applications.

7. CONCLUSIONS

This study presents a flexible multi-stage ensemble optimization (MSEO) strategy for articulated 2D-3D registration in image-guided wrist surgery. By leveraging direct optimization over machine learning, we are able to register a 3D wrist model to 2D intra-operative images without the need for a large database of labeled training examples. By employing a modular design incorporating differentiable rendering, ensemble optimization, and a multi-stage approach, MSEO demonstrates clear improvements over existing techniques on both synthetic and real datasets. These results highlight the potential of MSEO to address challenges in alignment accuracy through its superior performance across multiple evaluation metrics.

Despite its strengths, the method does not yet achieve the real-time, sub-millimeter accuracy required for precise visualization of internal bone structures in surgical environments. Challenges such as initialization sensitivity, skinning limitations in the 3D wrist model, and computational overhead from iterative rendering remain areas for improvement. Additionally, discrepancies between the silhouettes generated by soft rasterization and those obtained through segmentation introduce potential inaccuracies in the optimization process. Incorporating auxiliary data sources, such as real-time ultrasound tracking, could refine pose estimation by accounting for subtle bone movements. Improvements to the geometric accuracy of the 3D wrist modeling, and the development of advanced initialization strategies, could further address current limitations. Finally, exploring hybrid stochastic gradient-based optimization approaches to may further reduce sensitivity to local minima, while optimizing the rendering process for real-time performance could make MSEO more suitable for surgical applications.

In conclusion, while further refinement is necessary, the proposed MSEO framework provides a solid foundation for advancing 2D-3D registration in medical imaging. Its adaptability, modularity, and demonstrated accuracy position it as a promising tool for improving surgical guidance and for broadening the scope of medical image registration applications.

ACKNOWLEDGMENTS

We express our sincere gratitude to the AZ Monica hospital (Antwerp Campus) for facilitating the recording of the data and offering their support throughout the research process. This work was funded as part of the

HoloWrist project (https://holowrist.be/) under FWO grant number G0A8721N.

REFERENCES

[1] Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., and Black, M. J., "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in [*Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*], 561–578, Springer (2016).

[2] Wu, Z., Jiang, W., and Yu, H., "Analytical derivatives for differentiable renderer: 3d pose estimation by silhouette consistency," *Journal of Visual Communication and Image Representation* **73**, 102960 (2020).

[3] Dwivedi, S. K., Athanasiou, N., Kocabas, M., and Black, M. J., "Learning to regress bodies from images using differentiable semantic rendering," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 11250–11259 (2021).

[4] Kanazawa, A., Black, M. J., Jacobs, D. W., and Malik, J., "End-to-end recovery of human shape and pose," in [*Proceedings of the IEEE conference on computer vision and pattern recognition*], 7122–7131 (2018).

[5] Khaleghi, L., Sepas-Moghaddam, A., Marshall, J., and Etemad, A., "Multiview video-based 3-d hand pose estimation," *IEEE Transactions on Artificial Intelligence* **4**(4), 896–909 (2022).

[6] Tian, Y., Zhang, H., Liu, Y., and Wang, L., "Recovering 3d human mesh from monocular images: A survey," *IEEE transactions on pattern analysis and machine intelligence* **45**(12), 15406–15425 (2023).

[7] run Xiao, Z. and Xiong, G., "Computer-assisted surgery for scaphoid fracture," *Current Medical Science* **38**, 941–948 (2018).

[8] Liu, S., Li, T., Chen, W., and Li, H., "Soft rasterizer: A differentiable renderer for image-based 3d reasoning," in [*Proceedings of the IEEE/CVF international conference on computer vision*], 7708–7717 (2019).

[9] Song, J., Chen, X., and Hilliges, O., "Human body model fitting by learned gradient descent," in [*European Conference on Computer Vision*], 744–760, Springer (2020).

[10] Bălan, A. O. and Black, M. J., "The naked truth: Estimating body shape under clothing," in [*Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part II 10*], 15–29, Springer (2008).

[11] Tremblay, J., Wen, B., Blukis, V., Sundaralingam, B., Tyree, S., and Birchfield, S., "Diff-dope: Differentiable deep object pose estimation," (2023).

[12] Kavan, L., Collins, S., Žára, J., and O'Sullivan, C., "Skinning with dual quaternions," in [*Proceedings of the 2007 symposium on Interactive 3D graphics and games*], 39–46 (2007).

[13] Wu, G., Van der Helm, F. C., Veeger, H. D., Makhsous, M., Van Roy, P., Anglin, C., Nagels, J., Karduna, A. R., McQuade, K., Wang, X., et al., "Isb recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—part ii: shoulder, elbow, wrist and hand," *Journal of biomechanics* **38**(5), 981–992 (2005).

[14] Eschweiler, J., Li, J., Quack, V., Rath, B., Baroncini, A., Hildebrand, F., and Migliorini, F., "Anatomy, biomechanics, and loads of the wrist joint," *Life* **12**(2), 188 (2022).

[15] Allebosch, G., Vanhees, M., Luong, H., Veelaert, P., and Booth, B., "Real-time video enhancement for the removal of surgical lighting artifacts in computer-assisted orthopedic surgery," in [*21st IEEE International Symposium on Biomedical Imaging (ISBI 2024)*], IEEE (2024).

[16] Wang, C.-Y., Yeh, I.-H., and Mark Liao, H.-Y., "Yolov9: Learning what you want to learn using programmable gradient information," in [*European Conference on Computer Vision*], 1–21, Springer (2025).

[17] Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., and Sivic, J., "Megapose: 6d pose estimation of novel objects via render & compare," in [*CoRL 2022-Conference on Robot Learning*], (2022).

[18] Park, K., Mousavian, A., Xiang, Y., and Fox, D., "Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation," in [*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*], 10710–10719 (2020).

[19] Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., and Blaschko, M. B., "Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*], 92–100, Springer (2019).

[20] Beck, A. and Tetruashvili, L., "On the convergence of block coordinate descent type methods," *SIAM journal on Optimization* **23**(4), 2037–2060 (2013).

[21] Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.-Y., Johnson, J., and Gkioxari, G., "Accelerating 3d deep learning with pytorch3d," (2020).

[22] Mohammadi, Z. and Keyvanpour, M. R., "Similarity measures in medical image registration a review article," in [*2021 12th International Conference on Information and Knowledge Technology (IKT)*], 89–95, IEEE (2021).

[23] Yaegashi, Y., Tateoka, K., Fujimoto, K., Nakazawa, T., Nakata, A., Saito, Y., Abe, T., Yano, M., and Sakata, K., "Assessment of similarity measures for accurate deformable image registration," *J. Nucl. Med. Radiat. Ther* **3**(04) (2012).

[24] Prieto, M. S. and Allen, A. R., "A similarity metric for edge images," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(10), 1265–1273 (2003).

[25] Fan, S., Bao, Z., Dong, C., Liang, H., Xu, X., and Zhang, P., "Semantic similarity score for measuring visual similarity at semantic level," (2024).

[26] Rostamikhanghahi, H., Ingram, M., Booth, B., and D'Hooge, J., "Design of a linear flexible ultrasound array transducer for real-time tracking of the scaphoid during percutaneous scaphoid fixation," in [*2023 IEEE International Ultrasonics Symposium (IUS)*], 1–3, IEEE (sep 2023).

[27] Danckaers, F., Houtte, J. V., Booth, B. G., Verstreken, F., and Sijbers, J., "Statistical shape and pose model of the forearm for custom splint design," in [*2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*], 1669–1672 (2021).

[28] Van Houtte, J., Stanković, K., Booth, B. G., Danckaers, F., Bertrand, V., Verstreken, F., Sijbers, J., and Huysmans, T., "An articulating statistical shape model of the human hand," in [*Advances in Human Factors in Simulation and Modeling*], Cassenti, D. N., ed., 433–445, Springer International Publishing, Cham (2019).