Computers & Graphics 128 (2025) 104200

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

Accurate hand contact detection from RGB images via image-to-image translation ☆

Suzanne Sorli *, Marc Comino-Trinidad, Dan Casas [1]

Computer Science and Statistics Department, Universidad Rey Juan Carlos, Calle Tulipán s/n, 28933, Móstoles, Madrid, Spain

ARTICLE INFO

Keywords:
Computer vision
Hand tracking
Hand interaction

ABSTRACT

Hand tracking is a growing research field that can potentially provide a natural interface to interact with virtual environments. However, despite the impressive recent advances, the 3D tracking of two interacting hands from RGB video remains an open problem. While current methods are able to infer the 3D pose of two hands in interaction reasonably, residual errors in depth, shape, and pose estimation prevent the accurate detection of hand-to-hand contact. To mitigate these errors, in this paper, we propose an image-based data-driven method to estimate the contact in hand-to-hand interactions. Our method is built on top of 3D hand trackers that predict the articulated pose of two hands, enriching them with camera-space probability maps of contact points. To train our method, we first feed motion capture data of interacting hands into a physics-based hand simulator, and compute dense 3D contact points. We then render such contact maps from various viewpoints and create a dataset of pairs of pixel-to-surface hand images and their corresponding contact labels. Finally, we train an image-to-image network that learns to translate pixel-to-surface correspondences to contact maps. At inference time, we estimate pixel-to-surface correspondences using state-of-the-art hand tracking and then use our network to predict accurate hand-to-hand contact. We qualitatively and quantitatively validate our method in real-world data and demonstrate that our contact predictions are more accurate than state-of-the-art hand-tracking methods.

1. Introduction

Hand tracking is a rapidly growing research field that aims to estimate the 3D pose and shape of hands captured from the real world. One of the main motivations for solving this task is to provide a natural interface for human–computer interaction (HCI) problems. For instance, traditional interactions in Virtual Reality (VR) applications rely on hand-held controllers, which limit the visualization of hand movements. Hand tracking enables a more natural and intuitive interaction with virtual environments, allowing users to directly manipulate objects, perform gestures, and express themselves without physical controllers. Additionally, seeing and using one's own hands in VR enhances presence and immersion, making the experience more engaging and relatable. However, hand tracking is a highly challenging task due to natural self-occlusions, the similar appearance between the two hands, and the numerous degrees of freedom that must be estimated to recover hand motion". To tackle the grand challenge of tracking hands, many methods have been proposed. To simplify the problem, most of them focus on single-hand pose estimation, which has been successfully addressed using multi-camera setups [1,2] or depth cameras [3,4] to further simplify the task, although recent methods from RGB-only have also shown very accurate and robust results at tracking single hands [5–7]. However, we argue that the use of hands as a natural HCI interface requires methods that are capable of tracking the *two hands in interaction* from a single RGB camera. Unfortunately, very few works exist that tackle the two-hand tracking scenario [8–10] and, despite the promising results, they all share a common limitation: residual errors in depth, shape, or hand pose estimation prevent the accurate detection of hand-to-hand *contacts*.

To circumvent these shortcomings, in this paper, we propose an image-based method to estimate hand-to-hand contacts from a single RGB image. Our method builds on top of existing two-hand tracking solutions, enriching them with a camera-space probability map of hand contact. We argue that framing the hand contact problem on top of the 3D hand tracking problem provides many advantages: first, it enables the detection of contact even when 3D tracking is inaccurate, for example, due to errors in global pose estimation which produce 3D hands that do not touch in hand interaction motions; second, it makes

☆ This article was recommended for publication by A. Maciel.

* Corresponding author.
E-mail addresses: suzanne.sorli@urjc.es (S. Sorli), marc.comino@urjc.es (M. Comino-Trinidad), dan.casas@urjc.es (D. Casas).

[1] Work done prior joining Amazon.

https://doi.org/10.1016/j.cag.2025.104200
Received 17 July 2024; Received in revised form 26 February 2025; Accepted 7 March 2025
Available online 19 March 2025
0097-8493/

Fig. 1. Given a single RGB image of two hands in interaction, our model infers camera space 2D contact maps (in orange) that encode the pixels where the two hands are in contact. We demonstrate that our contact maps can be plugged into 3D hand tracking frameworks to improve accuracy.

our solution compatible with any existing two-hand tracking solution, for example, it can be plugged into either depth-based [11] and RGB-based [8] methods; and third, it can potentially be used as a new term in optimization-based methods for two-hand tracking, similar to other object-human contact terms that enforce soft constraints in the tracker [12].

We formulate our method as an image-to-image translation problem. Assuming a single RGB image as input, we first estimate pixel-to-surface correspondences [13] through a state-of-the-art 3D hand tracking method [9]. We then use a UNet architecture to translate the pixel-to-surface image into a probability contact map image. Since our network takes as input a pixel-to-surface encoding, it is agnostic to scene lighting conditions, shadows, and camera sensor modalities, which typically hinder the generalization of RGB methods.

To train our method, we propose a new pipeline to automatically annotate dense surface contacts in hand interaction sequences. To this end, we first use a commercial depth-based hand tracker to capture a collection of hand interaction sequences. We then fit a parametric surface hand model [14], and feed the resulting hand mesh sequences into a physics-based simulator [15]. This allows us to automatically detect and annotate per-vertex collisions, which we use to render ground truth contact maps for a variety of viewpoints. Additionally, for each frame, we also render ground truth pixel-to-surface correspondences automatically provided through the UV parameterization of the hand mesh. Finally, using a large dataset of pairs of pixel-to-surface images and contact maps, we train our UNet network. Notice that in this work we opt for not leveraging existing pretrained models for two reasons: first, pretrained models such as VGG [16] or ViT [17] are trained on very large datasets of RGB images obtained from real-world observations, however we design our pipeline to work non-photorealistic 4-channel images that store custom pixel-to-surface correspondences (i.e. DensePose [13]) and segmentation. We believe the significant difference in data distribution between our representation and natural images hinders the use of existing pretrained models as a useful prior: and second, we want to keep our model as light as possible. Recent Large Multimodal Models (e.g., LLaVA [18], CogVLM [19]) are capable of accurately parsing, understanding, and encoding real-world images, but require billions of parameters and large GPUs. In contrast, our method can run in desktop GPUs with less than 4 GB.

Through an exhaustive evaluation, we qualitatively and quantitatively demonstrate that our method is able to predict hand-to-hand contact labels with an accuracy higher than existing methods. In summary, our main contribution is twofold:

- To the best of our knowledge, the first method to explicitly learn to detect dense hand contact from RGB images, which can be plugged into any two-hand tracking framework.
- A pipeline to annotate dense per-vertex surface contacts in real-world sequences of two-hand interactions.

2. Related work

Single Hand 3D Tracking. The problem of tracking a *single* hand is nowadays a mature topic in the Computer Vision community. Pioneering works use multi-camera [20,21] or marker-based setups such as color gloves [22] to facilitate the detection of hand joints. Most of the early works use optimization-based [1,21,23] solutions that required solving an iterative process at each frame, which typically hinders interactive frame rates at runtime. To increase the robustness and performance of optimization-based methods, data-driven solutions such as random forests [1,3,24] or deep neural networks [25,26] have been proposed. These are used to guide or initialize the optimization process. More recently, leveraging the huge improvements in learning-based strategies enabled by Convolutional Neural Networks (CNNs), end-to-end single-hand methods for 3D pose estimation have appeared [5–7,27,28]. Self-supervised learning strategies have also been proposed by leveraging continuous hand motion information [29]. We also use a CNN in our method but, instead of estimating 3D hand joints, we estimate dense surface contacts.

Two-Hand 3D Tracking. Very few methods exist that tackle the challenging *two-hand* tracking problem. Oikonomidis et al. [30] introduced one of the first attempts, which uses a multi-camera setup to circumvent the challenges caused by the unavoidable strong self-collisions. A few follow-up works also use multi-view setups [31,32] and markers to ease the two-hand tracking problem, while others investigated the use of single depth sensors [11,33] which simplifies the setup but also provides sufficient cues to resolve depth ambiguities.

Closer to ours are the recent methods that consider a monocular RGB image as input. Wang et al. [8] introduce one of the first methods, which combined learned pixel-to-surface predictions with model-fitting to estimate the pose and shape of the two hands. Concurrently, Moon et al. [10] proposed a method that directly predicts 3D hand joint positions from RGB images. Follow-up works attempt to model the inherent depth ambiguity by explicitly modeling a visibility term [34], or by using probabilistic models for hand part segmentation [35] 3D pose [36]. Recent works [37] rely on coarse two-hand 3D keypoints as input to resolve such depth ambiguities. The method of Li et al. [9] uses a graph representation combined with attention modules to infer vertex positions of two interacting hands. Finally, the state-of-the-art Yu et al. [38] introduced ACR (Attention Collaboration-based Regressor), which reconstructs two interacting hands from a single RGB image by disentangling and reasoning the features of hands and parts. Despite the tremendous advances in 3D hand tracking of two hands from RGB, these methods suffer from residual errors in depth estimation that prevents the computation of accurate hand contacts.

Contact from RGB. Our work is also significantly related to the methods that, instead of just estimating hand poses, focus on estimating hand-object contact [39–42]. This problem has been addressed by data-driven pipelines that leverage large annotated datasets of hands manipulating rigid objects [43–45]. However, most of these methods estimate hand contact given 3D information of the scene or the target object [44,46]. Instead, we attempt to estimate camera-space contact labels for two hands without using any additional cue or annotation.

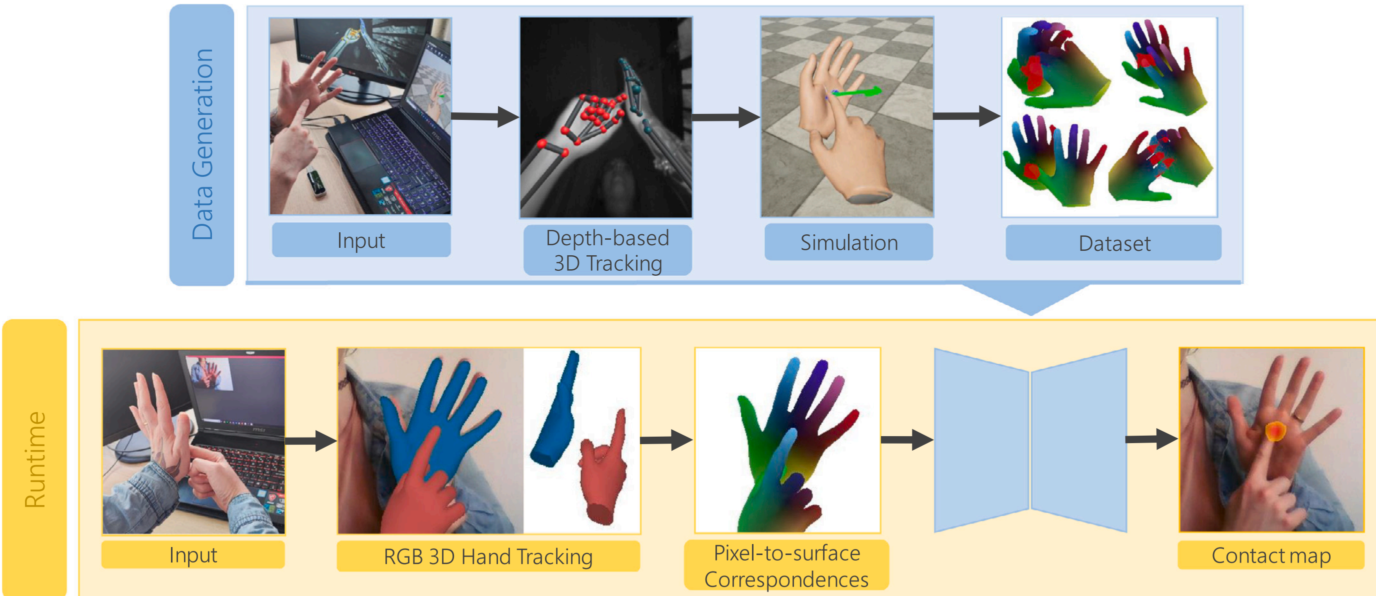

Fig. 2. Overview of our method to detect hand-to-hand interaction from RGB images. We first propose a pipeline to automatically annotate hand-to-hand contacts in sequences captured with a depth camera, which we use to create a dataset of pairs of surface-to-pixel correspondences and their contact map (top). At run time, we predict the 3D pose of each hand using a state-of-the-art method [9], but this often fails in capturing a global pose, which prevents hand contact estimation. We render the pixel-to-surface correspondences of the tracked hands and use our network to infer the accurate contact map (bottom).

Closest to us are the few methods that estimate contacts directly from RGB images. For example, Fieraru et al. [47] learn to estimate coarse subject-to-subject labels using a manually annotated dataset of human interactions. Huang et al. [48] are able to predict dense per-vertex contact labels by exploiting image information, without reconstructing 3D poses or 3D bodies.

3. Method

Our primary goal is to predict dense contact labels to augment the information inferred by a 3D hand pose estimation method when reconstructing two interacting hands from in-the-wild RGB images. To this end, we propose to use a strategy based on a popular image-to-image translation network [49]. Importantly, instead of working directly on RGB images, we first convert them into pixel-to-surface correspondences. This helps to circumvent many challenges related to in-the-wild imagery (e.g., lighting, shadows, background, etc.). Therefore, our system learns to convert estimated pixel-to-surface correspondences into contact probability maps.

Fig. 2 presents the different steps of our approach, which we describe in detail in the rest of this section. We first construct a dataset (Section 3.1) by simulating a wide variety of interaction sequences using a state-of-the-art hand simulator [15], and subsequently train our model (Section 3.2). At inference time (Section 3.3), we use a state-of-the-art method [9] to reconstruct the pose and shape of both hands but, as will be shown later, it sometimes fails at providing 3D hand poses that are in contact due to the depth ambiguity. We therefore render the hand poses using their color-coded dense correspondences and feed them into the network to infer the corresponding contact probability maps. Finally, we demonstrate that our contact maps enable us to refine the estimated hand poses by state-of-the-art hand tracking [9] to improve the 3D contact accuracy (Section 3.4).

3.1. Dataset generation

We require a dataset containing several sequences with two interacting hands and their corresponding contact labels. To the best of our knowledge, existing datasets in the literature, such as MSRA [50], which primarily focuses on single-hand tracking, or ARCTIC [51], which primarily focuses on hand-object interactions, do not fit our needs. Therefore, we first capture hand motion data using a commercial depth sensor [52] and then fit the MANO [14] parametric surface model to the captured motions to generate sequences of hand meshes in interaction. Following, we feed these meshes into the physics-based hand simulator CLAP [15], which was extended by Mueller et al. [11] to handle pairs of hands. This finally allows us to compute per-vertex contact labels. We captured 26 sequences to which we fitted the 3D hand models and performed physics-based simulations. Subsequently, each sequence was rendered at 30 frames per second (FPS), resulting in between 1800 and 3600 frames per sequence. This amounted to a total of 3.8×10^5 frames, all rendered at a resolution of 256 × 256 pixels. Additionally, at train time, we augment the data by generating 10 random rotations ranging between −100 and 100 degrees. For each frame, we generated two kinds of information: a pixel-to-surface map, which will be used as input to the network; and a binary mask with ones at the pixels where we have detected contact, which will be used as the prediction target. We use the pixel-to-surface map $\mathcal{N} : \Omega \rightarrow [0,1]^4$ from Mueller et al. [11] that assigns each pixel in the image domain Ω a color $[0,1]^4$ where the first 3 channels encode the dense correspondence to the hand surface and the last channel serves as a segmentation mask where 0 indicates the right hand, 0.5 the left hand and 1 indicates that the pixel does not correspond to any hand. Fig. 3 depicts some samples of our training set, overlaying the contact map labels on top of the pixel-to-surface image.

3.2. Network details

We learn to generate a binary mask that highlights regions with inter-hand contact in image space. For this task, we use an adapted version of the image-to-image translation network UNet [49], an architecture consisting of an encode-decoder scheme with skip connections. We take the Pytorch implementation by Buda et al. [53] as a reference. This uses 32 initial feature channels, which are doubled at each of the four levels on the multi-resolution pyramid. Each convolutional block consists of two 3 × 3 convolutions+activation followed by a max-pooling (left side of the pyramid) or upsampling (right side of the pyramid) units. We also use LeakyReLUs with a negative slope set of 0.1 and upsampling deconvolutions with bilinear upsampling units.

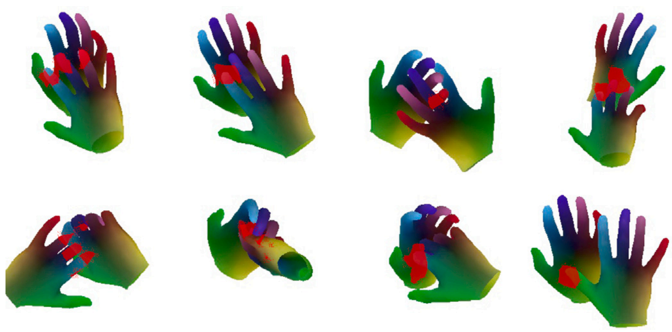

Fig. 3. Sample frames from the simulated sequences of our train set, with overlay ground truth contact map in red. Our dataset includes a wide variety of hand poses and interactions.

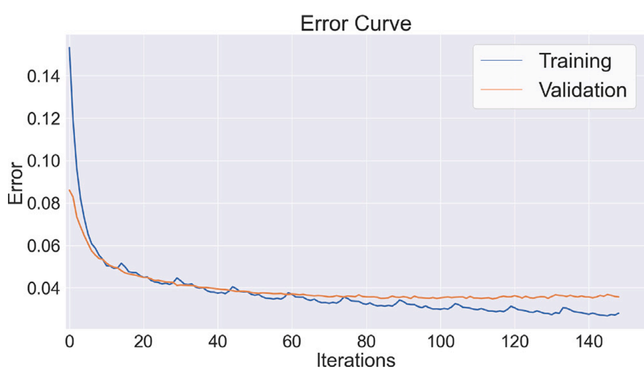

Fig. 4. Error values for our train and validation sets during training. An iteration in the horizontal axis corresponds to 250 batches of 8 images.

Our network takes a 4-channel image as input, consisting of the encoded hand dense-correspondences (*i.e.,* the pixel-to-surface encoding) and the hand segmentation mask, and predicts a contact probability map. To train our network, we use the standard binary cross-entropy loss to compare against the ground-truth contact maps together with an Adam optimizer, with a learning rate 10^{-4} and a batch size of 8. Fig. 4 depicts the loss curve while training, for train and validation sets.

3.3. Inference

At inference time, we can use any commercial RGB camera to capture sequences of an individual performing interactions with his or her hands. To provide an initial estimation of the hand tracking, we can use any state-of-the-art method for simultaneous two-hand tracking on the frames from the input sequences (e.g. [9]), and then convert the tracked sequences to MANO [14] representation. We render the corresponding hand meshes, from the same camera viewpoint, using a texture that encodes the surface-to-pixel color-coded correspondences. The rendered images are fed into the trained network which yields the desired contact maps.

3.4. Application: Global 3D hand pose optimization

Our inferred contact maps open the door to many downstream tasks in the area of hand tracking and human–computer interaction, where detecting accurate contact is crucial. To showcase a potential approach, we have used inferred maps to improve the accuracy of state-of-the-art 3D hand tracking methods with respect to 3D contact detection. To this end, starting from the 3D hand poses estimated with Li et al. [9], we optimize the global 3D position (*e.g.*, 3 DOFs) of the hand closest to the camera such that the vertices within the areas where contact is detected (in camera space) are in contact in 3D space. This effectively refines the hand positions estimated with state-of-the-art methods, estimating 3D hand poses that are in actual contact.

Results of this downstream task are showcased in the supplementary video and Fig. 8 (right), where we demonstrate the effectiveness of our method for optimizing the global 3D pose of hands tracked in real-world sequences.

4. Evaluation and results

We qualitatively and quantitatively evaluate our method in both synthetic scenes and real scenes. For each scene, we compare the detected two-hands contact using our method and the state-of-the-art method of Li et al. [9]. We also implement a straightforward RGB-only baseline consisting of a UNet network that directly predicts camera space contacts from RGB (*i.e.*, without using pixel-to-surface estimation). As a contact metric, we use True Positive Rate (TPR), True Negative Rate (TNR), and ROC curve. A pixel is labeled as a contact if the predicted probability is higher than a threshold τ_m. We consider a true positive a pixel that is labeled as contact and is predicted as contact; a true negative a pixel that is predicted as no contact and is labeled as no contact; a false positive a pixel that is predicted as contact and is labeled as non-contact; and a false negative a pixel that is predicted as non-contact but it is labeled as a contact.

4.1. Evaluation on synthetic data

To evaluate our method on synthetic data, we need a collection of hand interaction sequences with ground truth contact annotations and photorealistic 3D renders. To this end, we first capture data using Leap Motion [52] hand tracker and then convert the tracked skeletons into MANO [14] mesh representation. We then feed hand meshes into a physics-based simulator [15] that automatically computes ground truth contact maps. Finally, we also render the hand meshes using a photorealistic texture and background images. Fig. 5 (left) presents a few frames of a synthetic test sequence, with ground truth contact labels overlayed on top.

To test our method with these synthetic sequences, we first estimate the 3D hand pose of the two hands [9] and render the reconstructed meshes using a pixel-to-surface texture map. We then feed the renders into our network to estimate contacts. Fig. 5 shows representative frames of our results and compares them with ground truth contacts, the RGB-only baseline, and contacts computed using the output of the state-of-the-art 3D hand tracking method by Li et al. [9]. As it is obvious from the side view renders, the method of Li et al. can fail at providing a reliable 3D global pose, which precludes the computation of hand contact based on the regressed 3D information. Also, as can be seen in the figure, the RGB-only baseline mostly predicts noisy or weak contact signals. In contrast, our image-based method is capable of estimating dense contact maps that match the input image. Finally, the right column shows the result of optimizing the global hand pose using the inferred contact maps, which effectively computes 3D hand poses that are in actual contact.

We also conducted a quantitative evaluation of our method assuming ground truth 3D hand poses. This is an important test that allows us to disentangle the errors of our approach from the errors of the 3D hand-tracking method that we build upon. To this end, we render ground truth pixel-to-surface sequences of simulated hand interacting sequences, and the corresponding segmentation map, and pass it to our network. We then compute the F1 score of the output contact maps and compare it to the ground truth labels. Fig. 7 depicts this evaluation over 600 frames of the test sequence, demonstrating consistently high F1 values that highlight the precision and robustness of our method.

Finally, in Fig. 6 we show representative *test* frames of our dataset. This demonstrates that our approach generalizes to unseen hand poses at test time.

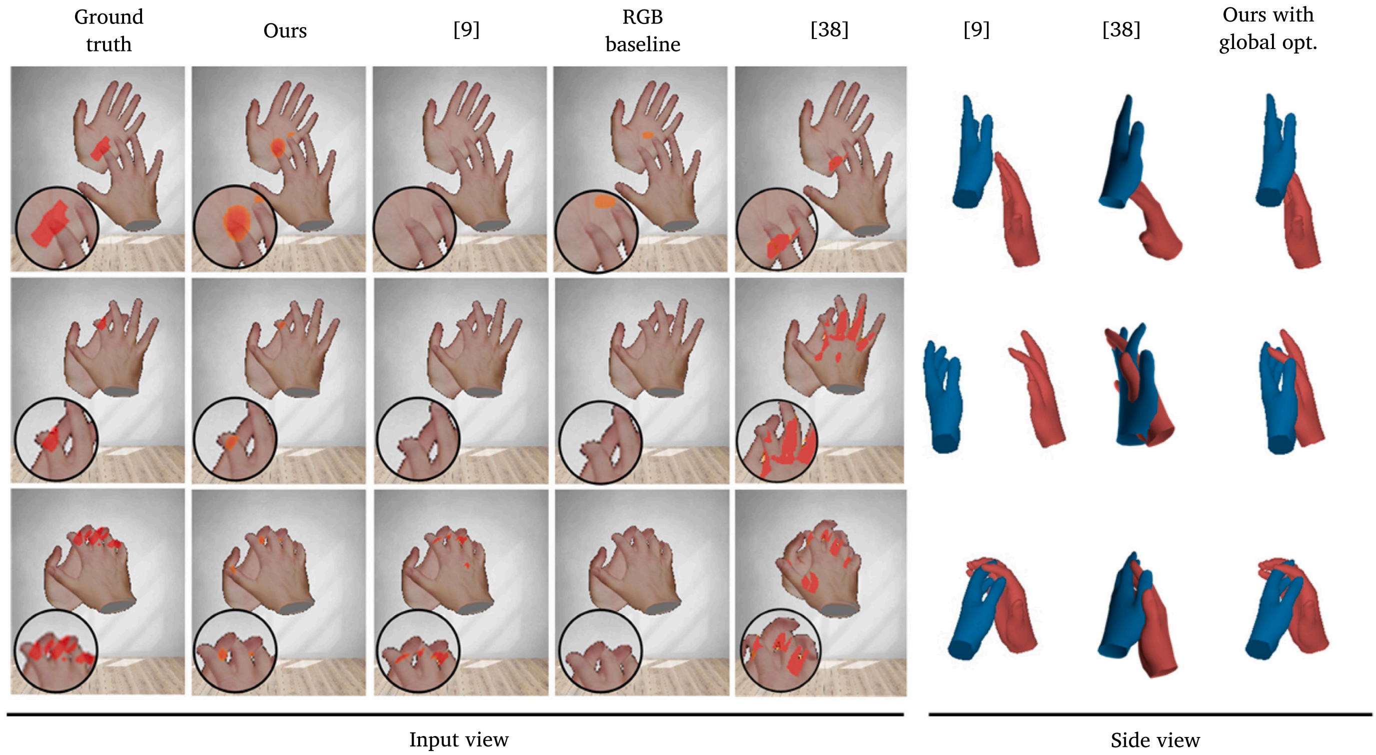

Fig. 5. Comparison to [9,38] and the RGB baseline on synthetic data, each frame with a zoom-in inset to contact area. Global 3D errors in [9,38] (6th and 7th columns, notice colliding or too separated hands) prevent the detection of contacts using the method of [9] (3rd column) and [38] (5th column). In contrast, our method (2nd column) closely matches the ground truth contacts (1st column, automatically annotated using simulation [15]), which can be used to optimize the global 3D position (8th column). Notice all colormaps were generated for values larger than 0.5 (for the RGB Baseline most values were under this threshold).

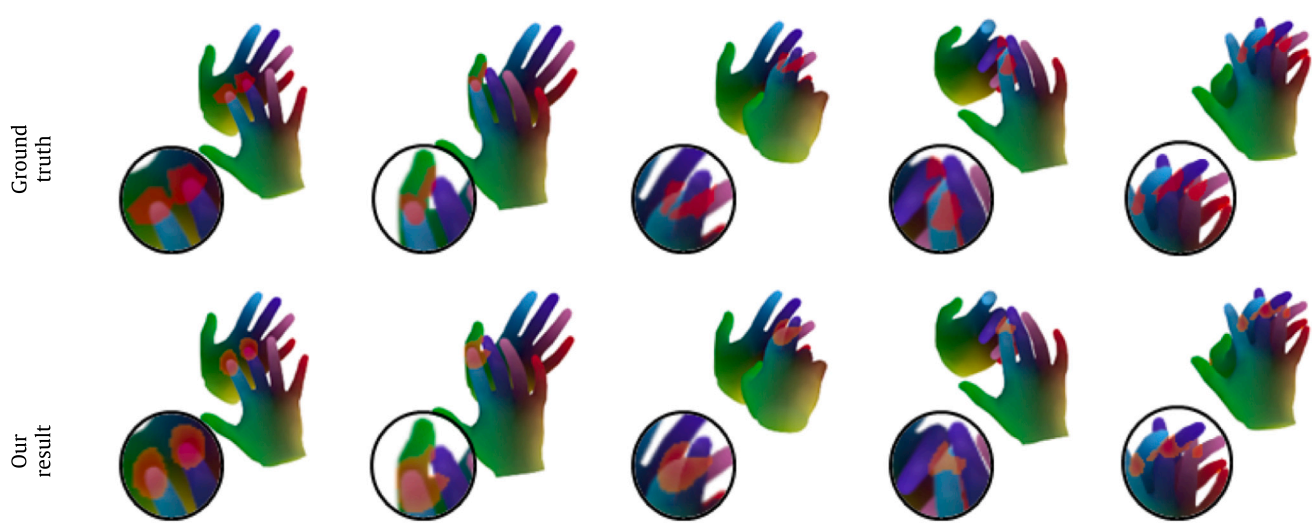

Fig. 6. To evaluate our approach, we show qualitative results on our test set. For this experiment, we use the ground truth pixel-to-surface image directly as input to our method (i.e., no hand tracking). This allows us to disentangle residual errors due to tracking issues. Results demonstrate that our predicted contact maps (bottom) closely match the ground truth (top).

4.2. Evaluation on real data

To showcase that our approach generalizes to unseen poses and natural images with uncontrolled lighting captured with various sensors, we quantitatively evaluate our results on real-world data captured from a desktop webcam and a mobile phone camera. To this end, we record hand sequences of interacting hands and manually annotate ground-truth contact maps using an interactive tool where users paint contact

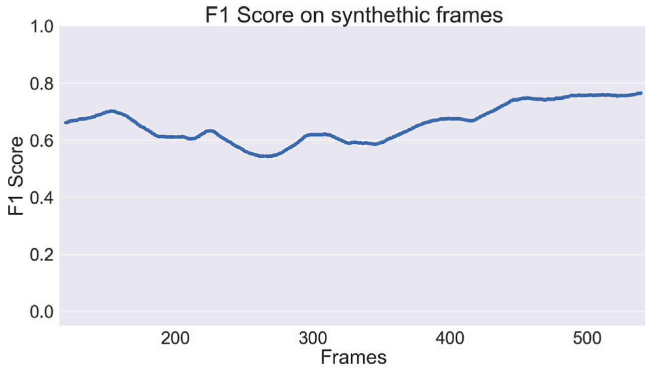

Fig. 7. Quantitative analysis of the predicted contact maps of our method on synthetic data.

on top of each frame. In total, we annotate 6540 frames captured with two different RGB sensors. Importantly, notice that none of the manual annotations were used to train our model, they are only used for evaluation purposes, since our method is solely trained on synthetic contact maps that are computed automatically using our physics-based hand simulator driven by motion captured data.

Fig. 8 presents a qualitative evaluation of representative frames of our method in a real-world test sequence. Results demonstrate that our predictions closely match the ground truth annotations, while the state-of-the-art method of Li et al. [9] and Yu et al. [38] cannot infer contact due to errors in global pose estimation. To highlight such errors, we show side views of the Li et al. [9] and Yu et al. [38] predictions, where it is very noticeable that hands are not in contact. In contrast, using our predicted camera-space contact maps enables the optimization of the global pose of the hand, yielding a correct 3D global pose estimation with accurate contact (right column). Additionally, for completeness, we show contact maps inferred with an RGB-only baseline (i.e., not using dense pose estimations as input to the network), which fails to predict accurate labels in most of the cases.

To evaluate our approach under a different lighting condition, we captured and annotated two additional real-world sequences, shown in Fig. 10, and report quantitative evaluations in Table 1 and Fig. 9. Our approach consistently overcomes the competing methods, yielding a True Positive Rate (TPR) 56% for our method, 8% for the method of Li et al. [9], and 1% with the RGB-only baseline, using a threshold of 0.40. Additionally, the ROC score for our method is 0.96, 0.55 for the method of Li et al. and 0.84 for the RGB-only baseline, which clearly highlights the superiority of our approach. Note that, for fairness in these quantitative analyses, we use thresholds different from 0.5, as otherwise, there were no predicted positive contacts for the RGB baseline. Thus, for each experiment, we selected the threshold closest to 0.5 that would produce some contact using the RGB baseline method. These results demonstrate that, regardless of the scene conditions, the contact accuracy of our method consistently outperforms the contacts detected using the 3D hand tracking method of Li et al. [9], Yu et al. [38], and the RGB-only baseline.

4.3. Qualitative results

For completeness, in the supplementary video and Figs. 1 and 10 we also show qualitative results of our method for a variety of subjects and sensors. Overall, our results demonstrate the success of our method in the challenging problem of estimating contact between two hands. To the best of our knowledge, our approach is the first to tackle this problem directly from RGB images.

Table 1
Quantitative evaluation of the sequence shown in Fig. 10 (middle).

	TPR ↑	TNR ↑	AUC ROC ↑	RMSE ↓
RGB baseline	1%	99%	0.84	0.028
Ours	**56%**	**99%**	**0.96**	**0.026**
Ours w/o seg.	38%	98%	0.92	0.036
Li *et al.* [9]	8%	99%	0.55	0.033
Yu *et al.* [38]	5%	99%	0.52	0.026

5. Conclusions, limitations, and future work

Hand tracking is a crucial feature for many mixed reality applications, enabling a more natural and direct interaction with the virtual environment. When simultaneously tracking both hands, the task becomes particularly ill-posed due to the numerous degrees of freedom that need to be estimated. To the best of our knowledge, we have presented the first image-based method for camera-space contact estimation of two interacting hands from a single RGB image. Our approach builds upon existing state-of-the-art two-hand tracking systems (e.g., [9]), which often suffer from residual errors that hinder the computation of hand contacts. We first render the estimated hand meshes using a pixel-to-surface texture and feed the renders to an image-to-image translation network, yielding dense contact labels. Our estimated contacts consistently achieve higher F1 scores than raw tracking systems in both synthetic and real scenes. Consequently, we can accurately estimate contact even when the underlying tracking fails. A key benefit of our method is that it is designed as a standalone component, not limited to a specific input sensor or modality. Since our network takes rendered pixel-to-surface correspondences as input, our method circumvents many challenges common in real-world images, such as illumination, sensor noise, or shadows. Finally, we plan to release the code of the method and the produced dataset upon acceptance of the work.

Limitations. As with most other learning-based methods, our system strongly relies on the quality of the training corpus. We believe we have produced a fairly rich set of training sequences and demonstrated the network's generalization capabilities. However, there are still some residual cases where we fail to predict correct contact labels. For instance, Fig. 11 shows some of our failure cases.

These failures can stem from two main sources. First, the dataset mainly consists of frames with both hands interacting, leading to very few frames with no interaction or only one hand. This can cause the model to hallucinate contact where there should not be any. Second, the variety of the training data impacts the model's predictions, such as the scale at which hands are captured. To address this, images are cropped to the region of interest around the hands to homogenize the scale. Additionally, we took measures such as altering hand shapes and aggressive data augmentation by rotation to enhance robustness. However, limitations remain; for instance, we limited the rotations to the image plane, whereas we could have also produced arbitrary 3D rotations since all our training data are 3D meshes that are later rendered to 2D.

Another limitation of our method is that it relies on the quality of the underlying tracker, to some extent. While we aim to solve the situations where the tracker cannot infer contact due to the depth ambiguities for the hands, our method will not work if the tracking is very unstable or estimates 3D poses that significantly differ from the observations.

Future work. We are currently considering multiple lines for future work. First, we want to investigate whether it is also possible to infer the forces involved in the interaction. Since we leverage a physics-based simulator to generate data, we can also generate labels for external

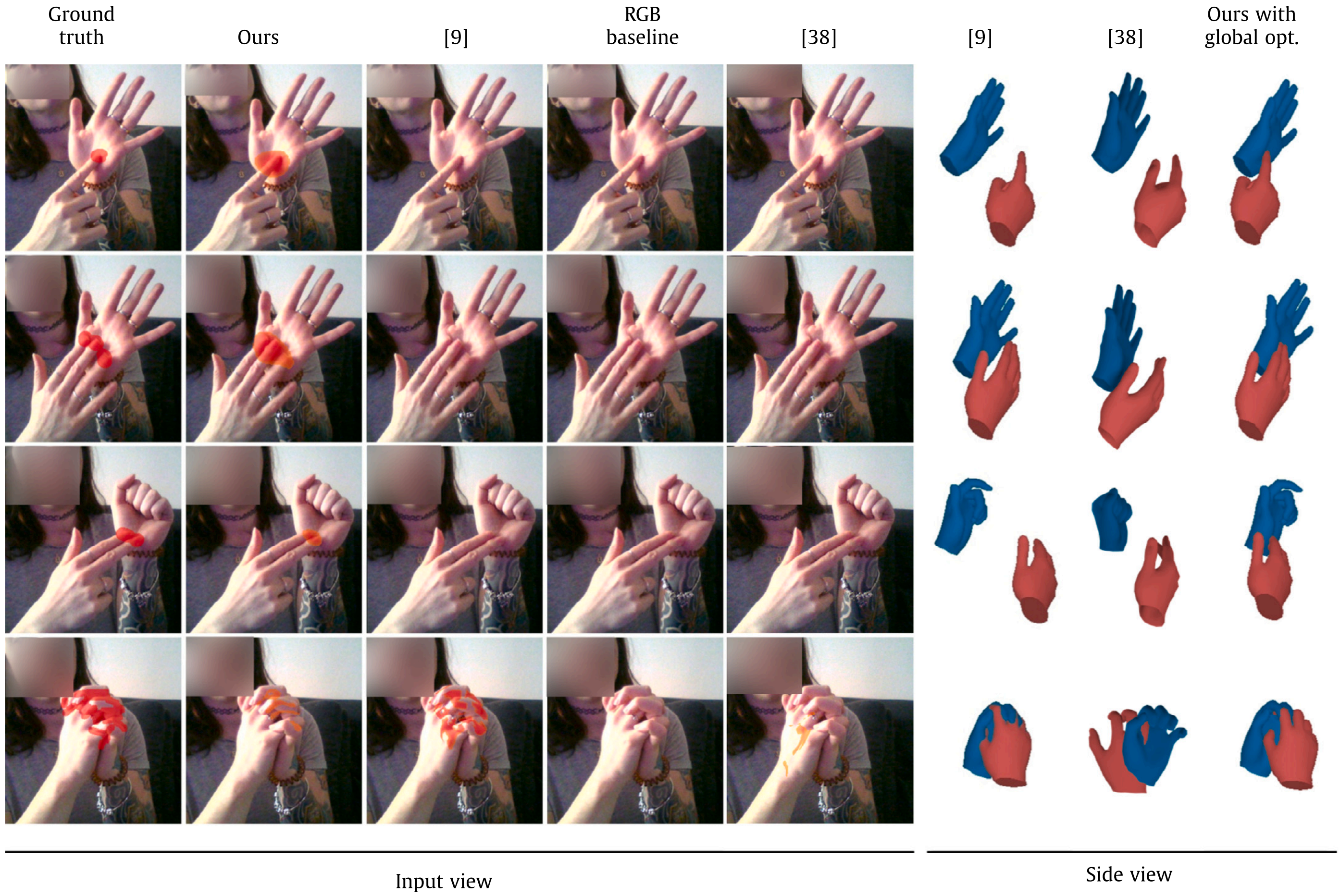

Fig. 8. Comparison to [9,38] on real-world data captured from a webcam. Errors in global 3D hand pose estimation (6th and 7th columns, notice colliding or too separated hands) prevent the detection of contacts using the methods of [9] (3rd column) and [38] (5th column). In contrast, our method (2nd column) closely matches the ground truth contacts (1st column, manually annotated), enabling the accurate estimation of 3D contacts by optimizing the global pose of the hand (8th column).

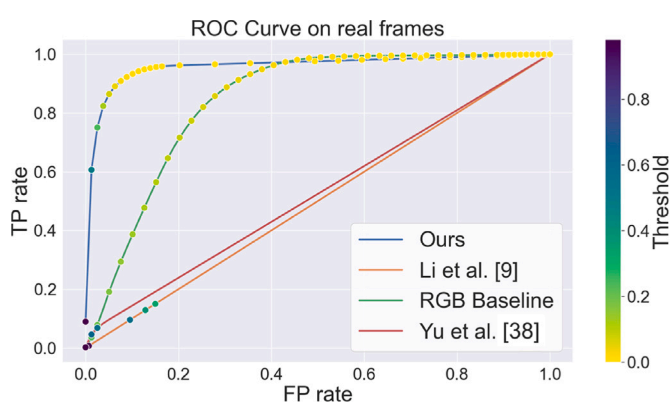

Fig. 9. Quantitative evaluation of the real sequence shown in Fig. 10 (center).

contact forces (*e.g.*, magnitude, direction) and we could try to learn them. Second, at the moment our pipeline runs at interactive rates (3–5 fps) but does not achieve real-time performance. Improving the performance of each of the steps of our system is required to speed up the system. Finally, our method heavily depends on the initial 3D poses estimated by the baseline state-of-the-art hand tracking method. Future research should look into relaxing this dependency.

CRediT authorship contribution statement

Suzanne Sorli: Software, Investigation, Conceptualization. **Marc Comino-Trinidad:** Writing – original draft, Supervision, Investigation, Conceptualization. **Dan Casas:** Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Dan Casas reports financial support was provided by Spanish State Agency of Research and the Spanish Ministry of Science and Innovation. Marc Comino-Trinidad reports financial support was provided by Rey Juan Carlos University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work has been partially funded by: the Spanish State Agency of Research under grant agreement TED2021-132003B-I00; by the Universidad Rey Juan Carlos through the Distinguished Researcher position INVESDIST-04 under the call from 17/12/2020; and by Spanish Ministry of Science and Innovation under grant agreement CNS2022-135996.

Fig. 10. Qualitative results of on real-world test sequences of different subjects and different lighting conditions. Our method predicts contact maps for a wide variety of hand-hand interactions directly from single RGB input.

Fig. 11. Failure cases of our approach. Our method sometimes gives false positives if only one hand is partially visible (left) or in very ambiguous hand interactions (center). Finally, our method sometimes struggle with heavily occluded and articulated hands (right).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2025.104200.

Data availability

Data will be made available on request.

References

[1] Ballan L, Taneja A, Gall J, Gool LV, Pollefeys M. Motion Capture of Hands in Action using Discriminative Salient Points. In: European conference on computer vision. 2012.

[2] Sridhar S, Oulasvirta A, Theobalt C. Interactive Markerless Articulated Hand Motion Tracking using RGB and Depth data. In: Proceedings of the IEEE/CVF international conference on computer vision. 2013, p. 2456–63.

[3] Sridhar S, Mueller F, Oulasvirta A, Theobalt C. Fast and Robust Hand Tracking Using Detection-Guided Optimization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2015.

[4] Malik J, Abdelaziz I, Elhayek A, Shimada S, Ali SA, Golyanik V, et al. HandVoxNet: Deep voxel-based network for 3D hand shape and pose estimation from a single depth map. In: Conference on computer vision and pattern recognition. 2020.

[5] Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, et al. GANerated hands for real-time 3D hand tracking from monocular RGB. In: IEEE/CVF conference on computer vision and pattern recognition. 2018, p. 49–59. http://dx.doi.org/10.1109/CVPR.2018.00013.

[6] Zimmermann C, Ceylan D, Yang J, Russell B, Argus M, Brox T. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In: International conference on computer vision. 2019.

[7] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. In: Conference on computer vision and pattern recognition. 2017, p. 4903–11.

[8] Wang J, Mueller F, Bernard F, Sorli S, Sotnychenko O, Qian N, et al. RGB2Hands: Real-time tracking of 3D hand interactions from monocular RGB video. ACM Trans Graph 2020;39(6).

[9] Li M, An L, Zhang H, Wu L, Chen F, Yu T, et al. Interacting attention graph for single image two-hand reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 2761–70.

[10] Moon G, Yu S-I, Wen H, Shiratori T, Lee KM. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In: European conference on computer vision. 2020.

[11] Mueller F, Davis M, Bernard F, Sotnychenko O, Verschoor M, Otaduy MA, et al. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. ACM TOG 2019;38(4):49.

[12] Sridhar S, Mueller F, Zollhöefer M, Casas D, Oulasvirta A, Theobalt C. Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In: European conference on computer vision. 2016.

[13] Alp Güler R, Neverova N, Kokkinos I. DensePose: Dense human pose estimation in the wild. In: Conference on computer vision and pattern recognition. 2018.

[14] Romero J, Tzionas D, Black MJ. Embodied hands: Modeling and capturing hands and bodies together. ACM Trans Graph 2017;36(6). http://dx.doi.org/10.1145/3130800.3130883.

[15] Verschoor M, Lobo D, Otaduy MA. Soft hand simulation for smooth and robust natural interaction. In: 2018 IEEE conference on virtual reality and 3D user interfaces. 2018, p. 183–90. http://dx.doi.org/10.1109/VR.2018.8447555.

[16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: International conference on learning representations. 2015.

[17] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International conference on learning representations. 2021.

[18] Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. In: Advances in neural information processing systems. 2023.

[19] Wang W, Lv Q, Yu W, Hong W, Qi J, Wang Y, et al. CogVLM: Visual expert for pretrained language models. 2023, arXiv:2311.03079.

[20] Wang R, Paris S, Popović J. 6D Hands: Markerless Hand-tracking for Computer Aided Design. In: Proceedings of the 24th annual ACM symposium on user interface software and technology. 2011, p. 549–58.

[21] Oikonomidis I, Kyriazis N, Argyros AA. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: International conference on computer vision. 2011, p. 2088–95.

[22] Wang RY, Popović J. Real-time hand-tracking with a color glove. ACM Trans Graph 2009;28(3). http://dx.doi.org/10.1145/1531326.1531369.

[23] Tagliasacchi A, Schröder M, Tkach A, Bouaziz S, Botsch M, Pauly M. Robust articulated-ICP for real-time hand tracking. Comput Graph Forum 2015;34(5):101–14. http://dx.doi.org/10.1111/cgf.12700.

[24] Sun X, Wei Y, Liang S, Tang X, Sun J. Cascaded Hand Pose Regression. In: Conference on computer vision and pattern recognition. 2015, p. 824–32.

[25] Sinha A, Choi C, Ramani K. Deephand: Robust hand pose estimation by completing a matrix imputed with deep features. In: Conference on computer vision and pattern recognition. 2016, p. 4150–8.

[26] Tompson J, Stein M, Lecun Y, Perlin K. Real-time continuous pose recovery of human hands using convolutional networks. ACM Trans Graph (ToG) 2014;33(5):1–10.

[27] Ge L, Ren Z, Li Y, Xue Z, Wang Y, Cai J, et al. 3D hand shape and pose estimation from a single rgb image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 10833–42.

[28] Boukhayma A, Bem Rd, Torr PH. 3D hand shape and pose from images in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 10843–52.

[29] Tu Z, Huang Z, Chen Y, Kang D, Bao L, et al. Consistent 3d hand reconstruction in video via self-supervised learning. IEEE Trans Pattern Anal Mach Intell 2023;45(8):9469–85.

[30] Oikonomidis I, Kyriazis N, Argyros AA. Tracking the articulated motion of two strongly interacting hands. In: Conference on computer vision and pattern recognition. IEEE; 2012, p. 1862–9.

[31] Simon T, Joo H, Matthews I, Sheikh Y. Hand keypoint detection in single images using multiview bootstrapping. In: Conference on computer vision and pattern recognition. 2017.

[32] Han S, Liu B, Wang R, Ye Y, Twigg CD, Kin K. Online optical marker-based hand tracking with deep labels. AMC TOG 2018;37(4):166.

[33] Taylor J, Tankovich V, Tang D, Keskin C, Kim D, Davidson P, et al. Articulated distance fields for ultra-fast tracking of hands interacting. ACM Trans Graph 2017.

[34] Kim DU, Kim KI, Baek S. End-to-end detection and pose estimation of two interacting hands. In: Proceedings of the IEEE/CVF international conference on computer vision. 2021, p. 11189–98.

[35] Fan Z, Spurr A, Kocabas M, Tang S, Black M, Hilliges O. Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-pixel Part Segmentation. In: 3DV. 2021.

[36] Wang J, Luvizon D, Mueller F, Bernard F, Kortylewski A, Casas D, et al. HandFlow: Quantifying view-dependent 3D ambiguity in two-hand reconstruction with normalizing flow. In: International symposium on vision, modeling, and visualization. 2022, p. 99–106. http://dx.doi.org/10.2312/vmv.20221209.

[37] Lee J, Sung M, Choi H, Kim T-K. Im2Hands: Learning attentive implicit representation of interacting two-hand shapes. In: Conference on computer vision and pattern recognition. 2023.

[38] Yu Z, Huang S, Chen F, Breckon TP, Wang J. ACR: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[39] Narasimhaswamy S, Nguyen T, Hoai M. Detecting hands and recognizing physical contact in the wild. In: Advances in neural information processing systems. 2020.

[40] Chen Y, Dwivedi SK, Black MJ, Tzionas D. Detecting Human-Object Contact in Images. In: Conference on computer vision and pattern recognition. 2023, URL https://hot.is.tue.mpg.de.

[41] He T, Gao L, Song J, Li Y-F. Exploiting Scene Graphs for Human-Object Interaction Detection. In: International conference on computer vision. 2021, p. 15984–93.

[42] Xie X, Bhatnagar BL, Pons-Moll G. CHORE: Contact, Human and Object REconstruction from a single RGB image. In: European conference on computer vision. Springer; 2022.

[43] Shan D, Geng J, Shu M, Fouhey D. Understanding human hands in contact at internet scale. In: Conference on computer vision and pattern recognition. 2020.

[44] Taheri O, Ghorbani N, Black MJ, Tzionas D. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In: European conference on computer vision. 2020.

[45] Liu S, Jiang H, Xu J, Liu S, Wang X. Semi-supervised 3d hand-object poses estimation with interactions in time. In: Conference on computer vision and pattern recognition. 2021, p. 14687–97.

[46] Cao Z, Radosavovic I, Kanazawa A, Malik J. Reconstructing hand-object interactions in the wild. In: Conference on computer vision and pattern recognition. 2021, p. 12417–26.

[47] Fieraru M, Zanfir M, Oneata E, Popa A-I, Olaru V, Sminchisescu C. Three-dimensional reconstruction of human interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[48] Huang C-HP, Yi H, Höschle M, Safroshkin M, Alexiadis T, Polikovsky S, et al. Capturing and inferring dense full-body human-scene contact. In: Conference on computer vision and pattern recognition. 2022, p. 13274–85.

[49] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015, p. 234–41.

[50] Qian C, Sun X, Wei Y, Tang X, Sun J. Realtime and robust hand tracking from depth. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2014, p. 1106–13.

[51] Fan Z, Taheri O, Tzionas D, Kocabas M, Kaufmann M, Black MJ, et al. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023, p. 12943–54.

[52] LeapMotion. 2016, https://developer.leapmotion.com/orion.

[53] Buda M, Saha A, Mazurowski MA. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. Comput Biol Med 2019;109. http://dx.doi.org/10.1016/j.compbiomed.2019.05.002.