Received: 2 July 2020 | Accepted: 15 July 2020

DOI: 10.1002/cav.1956

SPECIAL ISSUE PAPER

WILEY

Resolving hand-object occlusion for mixed reality with joint deep learning and model optimization

Qi Feng[1] | **Hubert P. H. Shum**[2] | **Shigeo Morishima**[3]

[1]Waseda University, Tokyo, Japan

[2]Northumbria University, Newcastle upon Tyne, UK

[3]Waseda Research Institute for Science and Engineering, Tokyo, Japan

Correspondence
Hubert P.H. Shum, Northumbria University, Newcastle upon Tyne, UK.
Email: hubert.shum@northumbria.ac.uk

Funding information
JSC ACCEL, Grant/Award Number: JPMJAC1602; JSPS KAKENHI, Grant/Award Numbers: JP17H06101, JP19H01129; JST-Mirai Program, Grant/Award Number: JP-MJMI19B2; Royal Society, Grant/Award Number: IES\R2\181024

Abstract

By overlaying virtual imagery onto the real world, mixed reality facilitates diverse applications and has drawn increasing attention. Enhancing physical in-hand objects with a virtual appearance is a key component for many applications that require users to interact with tools such as surgery simulations. However, due to complex hand articulations and severe hand-object occlusions, resolving occlusions in hand-object interactions is a challenging topic. Traditional tracking-based approaches are limited by strong ambiguities from occlusions and changing shapes, while reconstruction-based methods show a poor capability of handling dynamic scenes. In this article, we propose a novel real-time optimization system to resolve hand-object occlusions by spatially reconstructing the scene with estimated hand joints and masks. To acquire accurate results, we propose a joint learning process that shares information between two models and jointly estimates hand poses and semantic segmentation. To facilitate the joint learning system and improve its accuracy under occlusions, we propose an occlusion-aware RGB-D hand data set that mitigates the ambiguity through precise annotations and photorealistic appearance. Evaluations show more consistent overlays compared with literature, and a user study verifies a more realistic experience.

KEYWORDS

deep learning, hand tracking, mixed reality, occlusion, optimization

1 | INTRODUCTION

In recent years, mixed reality (MR) has drawn wide attention due to its versatile capabilities. Instead of rendering the whole scene from scratch like virtual reality, MR focuses on enhancing users' perception of the real world by overlaying virtual objects onto it.[1]

The hand is one of the key components in MR, and hand-object interactions are critical to a wide range of MR applications such as surgery simulations.[2] However, their practicality and immersive experiences are severely limited by occlusions. When we are holding some objects, it is very likely that fingers will partially occlude the object. If we render the virtual object with incorrect occlusions, an unrealistic "floating" illusion will lead to incorrect depth/distance perception, ruining an immersive MR experience.[3]

To resolve occlusions in MR, previous literature proposed various methods to resolve occlusions for virtual objects. However, the quality of their results is subpar when applied to hand-object interactions. Reconstruction-based methods utilize algorithms such as SLAM to obtain the model of the scene and render virtual objects through Z-buffer.[3] However, they are less viable for hand-object interactions that incorporate dynamic environments. The performance of tracking-based methods that track the object's contour through flow is limited due to the highly deformable hand shapes and severe occlusions.[4] Although depth-based methods can handle the aforementioned problems, sensor noise, misaligned boundaries, and unknown pixels within a close range limit their usefulness against egocentric scenes.[5]

In this article, we first proposed a photorealistic and occlusion-aware RGB-D hand data set that mitigates the ambiguity caused by occlusions and facilitate our learning-based system. By synthesizing pairwise occluded samples and augmenting them with realistic appearance, the data set containing precise annotations of hand poses and segmentation masks is generated with minimal manual input.

Making use of the generated data set, we propose an occlusion-aware joint learning system that shares information between the tasks of estimating hand poses and predicting semantic segmentation. By passing information between tasks, our system can predict more consistent results compared with existing single-task architectures. Taking advantage of the occlusion-aware data set, our joint-learning system is more robust in hand-object interactions. With precise estimations of hand joints and masks, the system facilitates the occlusion resolving task with our novel real-time optimization system.

Taking the advantage of occlusion-robust pose and mask information, we propose a novel real-time optimization system that renders correct occlusions when enhancing physical in-hand objects with a different virtual appearance. It overcomes the reconstruction-based approaches' limitation of not being able to handle dynamic scenes by efficiently reconstructing the spatial scene through a two-step optimizing-and-fitting method. By iteratively updating a parameterized hand model according to segmentation information and fit it back to tracked joints in real time, we calculate occlusion masks and augment virtual objects with correct representations.

Experimental results show high-quality overlays of augmented in-hand objects. A quantitative evaluation shows better performance over state-of-the-art approaches, a qualitative comparison shows better capabilities, and a user study verify more realistic MR experiences over literature. This research can be applied to egocentric applications that include hand-objects interactions such as tool-based simulations.

The contributions of this work are summarized as follow:

- A photorealistic and occlusion-aware RGB-D hand data set that facilitate occlusion-robust hand joints tracking and hand semantic segmentation. The data set is available for further research through the script: https://bit.ly/2TwCrS1
- An occlusion-aware deep-learning system that jointly estimates hand pose and semantic segmentation. With shared information between tasks, it can predict occluded hand-object interactions with high accuracy.
- A novel real-time optimization system for augmenting virtual objects that spatially reconstructs the scene through a two-step optimizing-and-fitting method.

The rest of the article is organized as follows. We revisit existing occlusion solutions and hand tracking methods in Section 2. In Section 4, we explain the proposed RGB-D data set and the joint learning process to estimate hand joints and masks. In Section 5, we present our novel two-step approach to resolve the occlusions. Implementation details, evaluations, and the user study are presented in Section 6. Finally, Section 7 concludes this work.

2 | RELATED WORKS

As the main goal of this work, we first revisit previous occlusion handling approaches in this section. Since the hand data set and the joint learning system are important components of this work, we also review previous learning-based methods and occlusion-aware data sets.

2.1 | Occlusions in mixed reality

A low-quality overlay in MR, such as incorrect occlusions, can easily break the immersive experience.[6] Methods in following literature estimate occlusions to correctly composite augmented objects with real environments without prior geometric knowledge.

Tracking-based solutions. Semiautomatic approaches, such as manually selecting the boundaries[7] and assigning the foreground and background objects,[4] estimate occlusions from the tracked contour of the selected foreground objects. These methods require manual input and implicitly assume that the objects have finite and constant boundaries. Moreover, the automatic contour extraction performance can be highly influenced by the insufficient resolution, false local minima, and errors in initialization.

Reconstruction-based solutions. Since the geometric relation between objects can be easily acquired if an accurate model of the real scene is provided, using a fast SLAM algorithm to reconstruct the environment is a popular option.[3] However, computationally expensive solutions require translating motions and textured environments, and have a major weakness of dynamic scenes with transient objects. Another type of approaches use predefined 3D models to perform fitting tasks via tracking.[8] The performance of these methods depends on the tracking robustness and is usually weak of deformable objects. Without an occlusion-robust tracking solution, these methods are prone to inconsistent results.

Depth-based solutions. Using additional depth sensors to obtain the per-pixel depth information directly can serve the purpose as well. However, temporal noise and misaligned depth edges and other underlying problems lead to low-quality results. Chao et al.[5] proposed an edge snapping algorithm to improve the consistency along the boundaries. However, most state-of-the-art algorithms are not practical when the computational cost is a problem that cannot be ignored in modern MR applications.[9] Refining the obtained depth in a "layered" fashion with cost-volume filtering[10] can achieve real-time performance, but it generalizes poorly for complex scenes such as interactions. Besides, the hand being simultaneously foreground and background object would make color-based segmentation impractical.

By leveraging the efficiency of tracking-based and model-based methods, our proposed real-time method solve occlusions in hand-object interactions without introducing a lengthy initialization, extra sensors, or an expensive process of reconstructing the entire scene.

2.2 | Occlusion-aware hand tracking

Learning-based approaches. Vision-based 3D hand pose prediction is a challenging task due to its high degree of freedom articulations and severe self and hand-object occlusions. Markerless approaches introduce generative components to improve the estimation between the simulation and the observation, such as consistencies between frames,[11] iterative closest point,[12] particle swarm optimization,[13] and so forth. However, most methods require a lengthy initialization process, and its accuracy highly depend on the quality of observations. To address these limitations, learning-based discriminative components have become a popular choice in recent years.[14] Although being beyond the scope of this study, adapting MANO[15] to solve interactions between hands,[16] and joint tracking of hand and object[17] are all promising directions for improvements. In this work, we use a learning-based approach to precisely and efficiently predict hand poses without manual initialization.

Occlusion-aware data sets. One of the major issues of learning-based hand tracking methods is difficulties in preparing training data with correct 3D annotations. Recently, a handful of high-quality data sets for 3D hand pose estimation are released.[13] Even data set constructed upon manual annotations exists,[14] inaccuracy and insufficient size are problems that cannot be ignored. Multiview approaches[18] suffer from the limitation of occlusions due to their outside-in setups. To obtain accurate paired data, some works render synthetic RGB-D hand-object images with hand models and virtual cameras.[14] However, existing CNNs-based approaches that are trained on synthetic data generalize poorly due to the domain gap between synthetic and real-world images. To improve the accuracy of occlusion-aware hand tracking from RGB-D input, our method leverages a CycleGAN and incorporates the geometric consistency loss[19] to synthesize a photorealistic RGB-D hand data set.

3 | THE FRAMEWORK OVERVIEW

To achieve the goal of augmenting in-hand objects with correct occlusions in real-time, our approach consists of two major components (see Figure 1). The first one is an occlusion-aware joint learning framework for (a) hand pose estimation and (b) semantic segmentation. This involves building an occlusion-aware hand database, a joint tracking module, and a segmentation module. The second one is a real-time optimization-based occlusion resolving system (c) for virtual object augmentation. This involves optimizing a hand model using the estimated segmentation mask, fitting the model with the tracked hand pose from the joint learning step, and resolving occlusion masks for augmenting virtual objects.

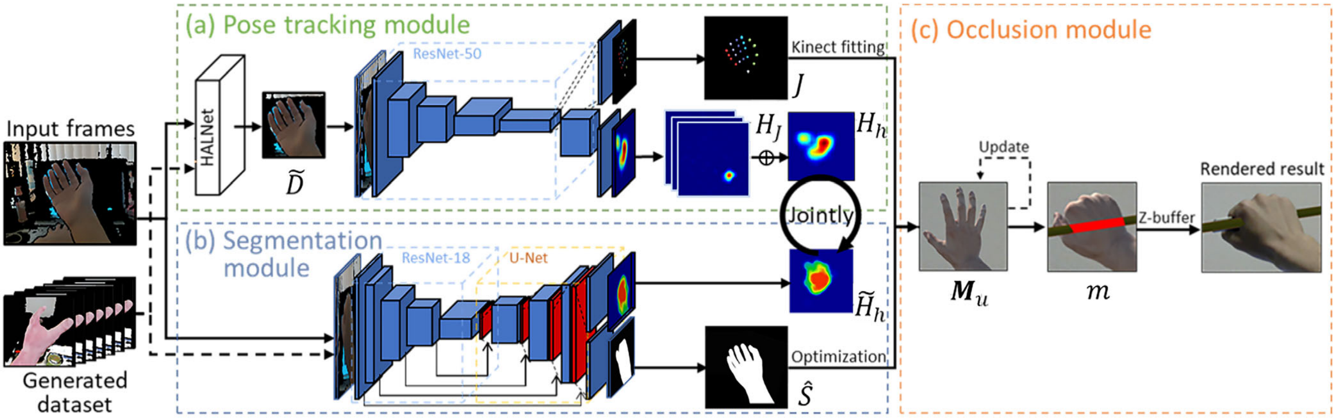


FIGURE 1 The architecture of our hand-object occlusion resolving system

4 | OCCLUSION-AWARE JOINT LEARNING FOR POSE TRACKING AND SEGMENTATION

4.1 | The photorealistic and occlusion-aware data set

We proposed a photorealistic and occlusion-aware RGB-D data set to facilitate learning-based pose tracking and semantic segmentation in hand-object interactions. This is motivated by the difficulty to annotate hand poses and segmentation in occluded scenes. Capture-and-annotate methods are limited due to ambiguities, glove-based methods yield different appearance and are not suitable for bare-hand applications. To acquire accurate pose information and segmentation masks for hand-object interactions, synthesizing samples and annotations is a more efficient and suitable way compared with capture-based methods.

To efficiently synthesize the photorealistic and occlusion-aware RGB-D data set, we repurpose an existing synthetic RGB-D hand data set.[14] It contains samples with hand-object interactions and hand joint annotations. To adapt it to our use, we first re-render the hand into binary masks for the semantic segmentation task. Inspired by Mueller et al.,[19] we then use the generated segmentation masks as geometric constraints to transfer photorealistic appearance to synthetic samples by training a CycleGAN. To ensure annotations stay correct after the transfer, we calculate the geometric consistency loss from the estimated and real silhouettes:

$$L_{geo} = -\sum_{i} (S_i log\widehat{S}_i + (1 - S_i)log(1 - \widehat{S}_i)), \tag{1}$$

where S is the rendered segmentation and $\widehat{S}$ is the mask of the generated sample. The pipeline is explained in Figures 2 and 3 shows some samples of the data set. We only show the synthetic-to-real half of components for simplicity.

By reducing the domain gap between synthesized and real-world scenarios, our approach improves the accuracy of learning-based approaches. A photorealistic RGB-D hand data set with occlusions that contains 40,000 precise joint and segmentation annotations is created to facilitate applications including hand tracking and semantic segmentation.

4.2 | Hand semantic segmentation and pose estimation under occlusions

With precise pose annotations and semantic segmentation of photorealistic hand data available, we propose a deep-learning system that jointly estimates hand semantic segmentation and tracks joints. As illustrated in Figure 1a,b, we pass the input to our joint-learning system with a resnet-structured pose tracking module and a U-net-structured segmentation module running parallelly to each other.

To achieve a more coherent estimation even under occlusions, we exploit the information of joint annotations and inform the other task of potential uncertainties with concatenated heatmaps of pose estimation. More specifically, in addition to predicting 3D coordinates of each joint, 2D Gaussian heatmaps of every joint is also created with the pose

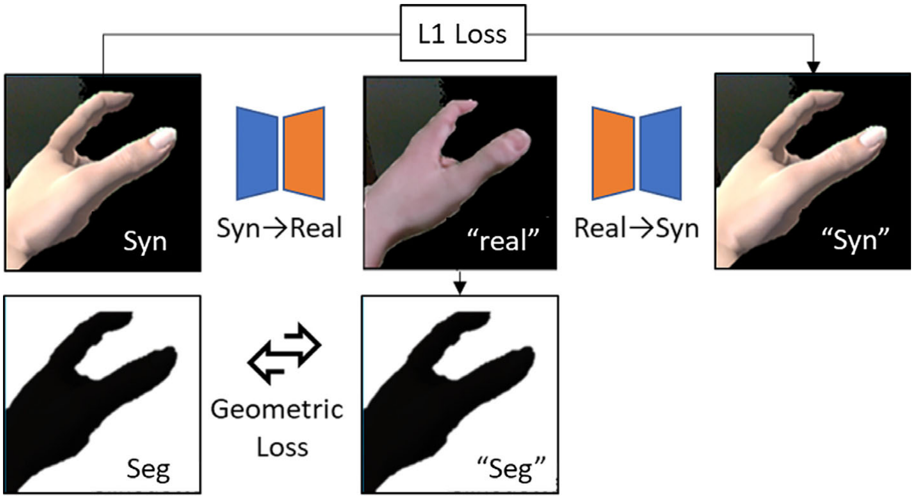


FIGURE 2 The CycleGAN architecture of our photorealistic RGB-D data generating network

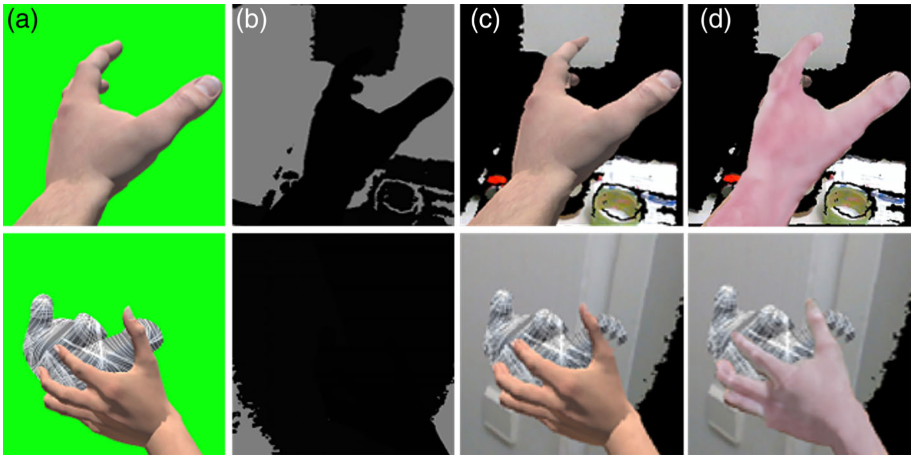


FIGURE 3 An example of generated photorealistic RGB-D hand data set. Images from left to right are (a) synthesized RGB; (b) synthesized Depth; (c) synthesized RGB-D; (d) photorealistic RGB-D

tracking module. We find that two tasks are complementary to each other since hand joints should always located within the hand contour, hence we concatenate and convey heatmaps to the semantic segmentation module. With a heatmap loss calculated to reduce false-positive predictions, our joint learning system has improved accuracy compared with two separate modules without communications.

4.2.1 | The pose tracking module

To estimate hand poses with improved accuracy and robustness, we take advantage of our generated photorealistic hand data set and propose the pose tracking module to predict hand poses. With the input of an RGB-D image, the pose tracking module is trained to regress 3D locations of 21 hand joints. As additional information to be shared with the other task, 2D Gaussian heatmaps are also output in image space with the pose estimation module.

A two-step localizing-and-tracking method is used to improve the robustness of the network. We adapt the HALNet[14] and trained with the proposed data set to localize the hand when an image is inputted. $\widetilde{D}$ that contains hand will be cropped from the input RGB-D frame D and passed to the next step. We then propose a pose estimation network bases on a modified ResNet-50 structure with reduced layers to achieve real-time performance. By minimizing the Euclidean loss between predicted joints and ground truth $\widehat{J}$, our hand pose tracking module can estimate 3D hand joints' coordination J in real-time during usage.

$$d_{pred} = \sum_{i=0}^{N} \|J - \widehat{J}\|_2^2. \tag{2}$$

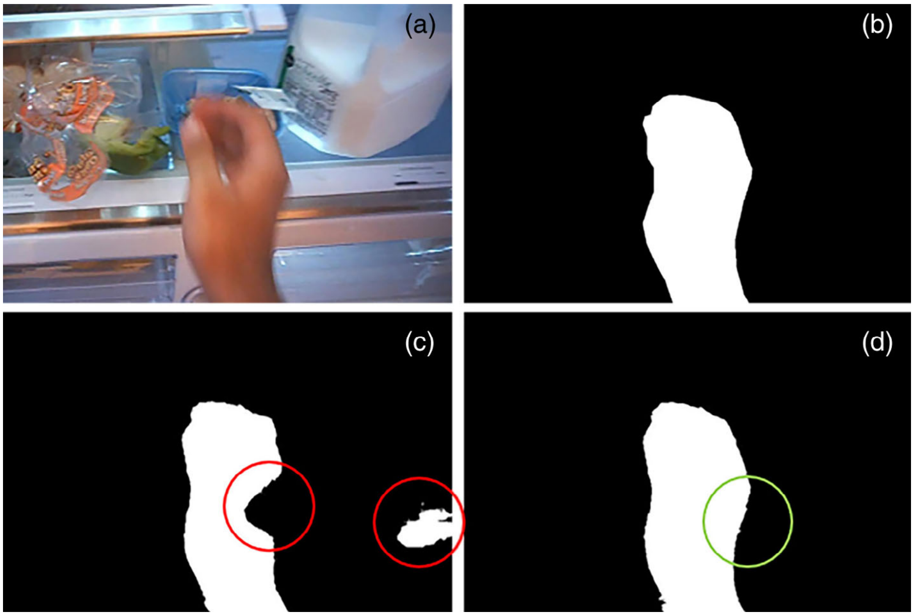


FIGURE 4 A comparison between segmentation masks estimated with and without heatmap loss. (a) Input color image. (b) Ground truth segmentation mask. (c) Estimated mask without the heatmap loss. (d) Estimated mask with the heatmap loss

Since hand pose annotations being a high-level information is expensive to acquire but highly correlated to and beneficial for different tasks (model reconstruction, normal estimation, etc.), we exploit the learned image to pose mapping through heatmap representations. As 2D likelihood heatmaps H_j are regressed during pose estimation for each joint, we concatenate heatmaps of each joint to obtain hand heatmap H_h, and pass the information to the segmentation module during training.

4.2.2 | The semantic segmentation module

To facilitate the real-time optimization system in the next step, we propose a semantic segmentation module to estimate hand masks from an image input. To take advantage of hand pose knowledge, the segmentation module outputs intermediate heatmaps for mask estimations, and calculate an additional heatmap loss to ensure that the hand joints fall within the segmentation estimation. Combined with the synthesized occlusion-aware pairwise images and masks, this module can handle occluded scenes with improved performance.

Structurewise, the segmentation module consists of a U-Net structure with the encoder part replaced with a ResNet-18 backbone. Considering the binary output mask, we choose the dice coefficient as our segmentation loss function.

$$L_{dice} = \frac{2|\widehat{S} \cap S|}{|\widehat{S} + S|}. \tag{3}$$

The $\widehat{S}$ in Equation (3) is the estimated segmentation, while S is the ground truth. The architecture of our hand semantic segmentation network is shown in Figure 1b.

To reduce false-positives in estimated masks, we leverage the information, heatmap of the hand H_h obtained from the pose tracking module. Apart from the main loss between the segmentation masks, with the average pooling, we create activation maps at the same time for calculating the complementary heatmap loss between the $\widetilde{H}_h$ obtained by the segmentation module and the H_h with Euclidean loss. The weight of heatmap loss is set at 0.1 during training, and the resolution is downscaled to 640 × 360 to maintain a stable speed. As demonstrated in Figure 4, we can clearly see the effectiveness of guiding the semantic segmentation task through heatmaps passed by the pose tracking module.

5 | VIRTUAL OBJECT AUGMENTATION WITH REAL-TIME OPTIMIZATION

In this section, we explain our novel real-time optimization system that resolves the occlusions in hand-object interactions through a two-step optimizing-and-fitting method. With the hand pose and semantic segmentation information available,

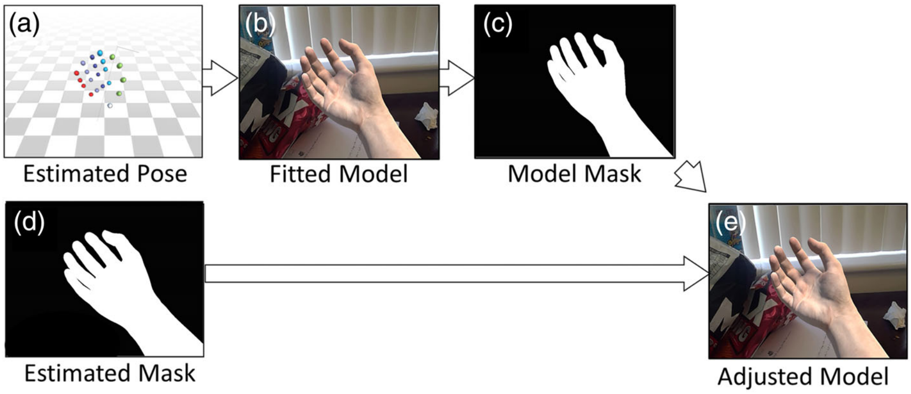


FIGURE 5 The process of updating the hand model during runtime with estimated hand poses and masks. The model is optimized by minimizing the distance between observed and model's rendered segmentation

we spatially reconstruct the region of interest with high accuracy by first iteratively optimizing a virtual hand model based on user's hand and then fitting the updated model to estimated hand poses.

This system design circumvents the limitations of previous occlusion resolving approaches effectively. For reconstruction-based methods, we overcome the constraint of only being able to recover a static scene by fitting the optimized model to estimated joints in real time. For tracking-based methods, the problem of low-quality outcomes against changing shapes or under severe occlusions is solved by our occlusion-aware joint learning system. Instead of tracking contour directly, we calculate it through more occlusion-robust hand poses. With local models of the hand and the virtual object available, we then augment the object through an occlusion mask calculated in real time.

5.1 | Hand model optimization

With the current frame of hands available, a hand pose and semantic segmentation are estimated with the joint learning system and inputted to this occlusion module to optimize a virtual hand model M_d in real time. This iterative process (Figure 5) is effective and efficient against hand-object interactions by only reconstructing models. More specifically, by fitting the (b) current model according to (a) the estimated hand pose J and projecting the model back to the image plane where the scene is rendered, we can obtain (c) a binary mask $\widetilde{S}$. At the same time, we can acquire (d) an estimated hand segmentation mask $\widehat{S}$ through our segmentation module. The hand model consists of fingerwise components and a palm component, and each has parameters of vertical and horizontal scale. We then update (e) the current model M_u to minimize the euclidean distance $d_S = \|\widehat{S} - \widetilde{S}\|$ between observed and rendered hand masks.

To further enhance the stability of outputs, we take consistency into consideration and minimize the distance through a step-based iterative optimization. The initial step for updating the scale of the model is 0.2 for every 30 frames. When the model meets a plateau for successive 300 frames, we upscale/downscale the step by 50%. Since we want to achieve a more stable output, the optimization is stopped when the step size goes smaller than 0.02 to save computational power and prevent flickering effects in the implemented real-time application.

5.2 | Hand model fitting and virtual objects augmentation

To cope with fine occluding edges between user's hand the in-hand objects, we propose a way to calculate occlusion masks through refined depth relations by comparing the reconstructed hands and objects to be augmented in a virtual environment. Existing depth-based methods suffer from problems including noise and misalignment, and their quality deteriorates when the distance from targets getting closer.

More specifically, we solve occlusions with the updated hand model by fitting it to the joints acquired through the pose tracking module in our joint learning framework. Our approach minimizes the fitting energy with regard to the optimized hand model. The updated hand model is displaced to minimize the distance d_j between the captured hand joints J_i and the current hand model $M_u(i)$:

$$d_j = \sqrt{\sum_{i=0}^{N} (d_J(i) - \widehat{d}_{M_u}(i))^2}, \tag{4}$$

where $d(i)$ is the normalized distance obtained by $d(i) = r_i/\sqrt{S}$. The r_i is the distance between the feature joint $J_i(i = 0,1, \dots ,N)$ and the root of the hand J_r. We fit the optimized hand model $\boldsymbol{M}_u$ back according to the acquired joint coordination J to calculate the occluding mask $\boldsymbol{m}$, the spatial location is shown on the image plane to which the invisible part of the virtual object $\boldsymbol{V}$ corresponds.

We decide the label of each point as "visible" or "invisible" of $\boldsymbol{V}$ based on the comparison between the 3D displacement of $\boldsymbol{V}$ and the optimized hand $\boldsymbol{M}_u$ to determine the occluding mask $\boldsymbol{m}$. During rendering, pixels of $\boldsymbol{V'}$ labeled with "invisible" will not be rendered to represent the occlusion. Based on the acquired occlusion mask, some portion of the virtual object model will be masked invisible, while other portions remain visible to the user. This process is done by frame and will remain robust even under strong motion.

6 | EVALUATIONS

6.1 | The mixed reality application

We implemented a complete table-top application (Figure 6) to showcase the idea, verify the quality of masks, and conduct a user study. This Unity3D application allows users to use their bare hands to interact with real objects augmented with virtual appearances. The frame rate was fixed at 30 fps with a resolution of 1,440 × 1,440 per eye using a PC with a Intel 7800X CPU and a NVIDIA RTX 2080Ti. Although a video-see-through equipment (Intel RealSense SR300) is used during the experiment, our system also works with optical-see-through devices.

6.2 | Qualitative results

To qualitatively verify the applicability of our proposed system when applied to hand-object interactions, we compared it to previous real-time occlusion solutions in the following five aspects. First, the system should be able to resolve the occlusion with a moving viewpoint. Restricting the viewpoint will significantly reduce the practicability. Second, the placement of in-scene objects will change constantly, and thus being able to handle dynamic scenes is critical. Moreover, additional equipment and complex implementations can limit usability. The detailed comparison is shown in Table 1. Our approach can handle dynamic scenes with moving objects and egocentric viewpoint in real-time with a simple setup.

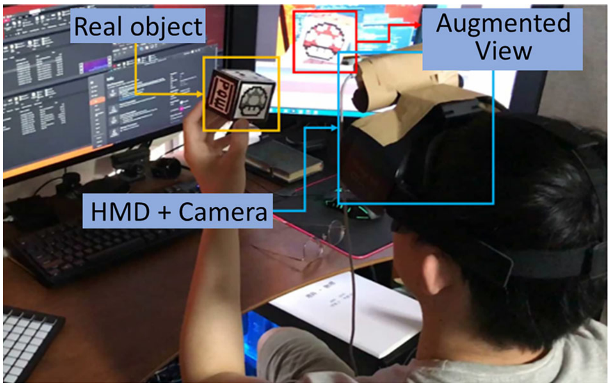


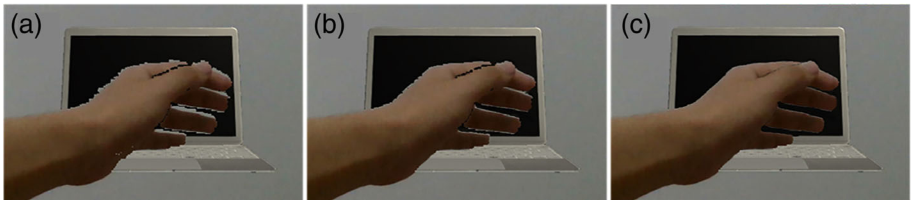
FIGURE 6 The configuration of the implemented application

TABLE 1 A comparison between the proposed method and the previous methods

Method	Viewpoint	Scene	Mutual occlusion	Equipment
Lu[23] Depth based	Restricted	Static	No	Stereo cameras
Tian et al.[4] Contour based	**Arbitrary**	Static	No	**RGB camera**
Dong et al.[24] Depth based	Restricted	**Dynamic**	**Yes**	TOF camera
Tian et al.[8] Reconstruction based	**Arbitrary**	Static	**Yes**	RGB-D camera
Holynski et al.[3] Reconstruction based	**Arbitrary**	Static	**Yes**	**RGB camera**
Walton et al.[10] Depth based	**Arbitrary**	**Dynamic**	No	RGB-D camera
Our method	**Arbitrary**	**Dynamic**	**Yes**	RGB-D camera

Note: Bold text in the table shows the best capability for most applications/run most efficiently/require minimal setup among proposed approaches.

FIGURE 7 A MR scene rendered with occlusions based on (a) naive approach that uses raw depth, (b) CVF occlusion,[10] and (c) our approach



We evaluate our system and verify this is a better method compared with the naive method that decides the visibility of each pixel based on the raw input of the RGB-D camera, the state-of-the-art CVF occlusion approach.[10] We exclude SLAM-based methods due to their unrealistic requirements of a stable and rigid environment in hand-object interactions. By placing virtual objects in the scene to interact with the scene geometry, we implemented a traditional AR scenario of object insertion to evaluate the accuracy of the occlusion mask and rendered object. Direct results of rendered virtual objects can be observed in Figure 7. The readers are also referred to the supplementary video for further results.

6.3 | Quantitative results

6.3.1 | Tracking under occlusions

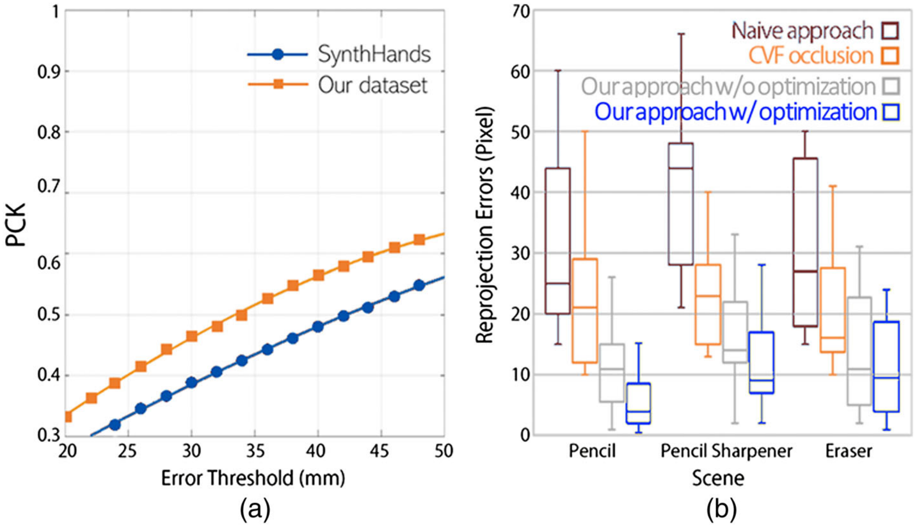
To verify the effectiveness of the improved photorealistic hand-object RGB-D data set, our model is trained with a similar architecture to the JORNet[14]

with Caffe framework. The weight of our network is initialized based on the original ResNet50 trained with ImageNet.[20] We use percentage of correct keypoints (PCK) as the measure to evaluate the accuracy of our approach. After training 45,000 iterations with the same configuration based on the original SynthHands[14] and our improved data set, we benchmark both approaches with the stereo tracking benchmark data set,[21] which consists of 12 sequences of paired RGB-D images. Figure 8a presents the result that and our approach outperforms the original method trained with synthetic data. With a threshold set at 50 mm, the accuracy is significantly improved from 0.55 to 0.63.

6.3.2 | Ablative analysis

To validate the quality of the overlay, we mainly focus on the reprojection error in pixels of the rendered objects. Since the egocentric HMD works differently from the traditional screens, the screens are positioned closer to the user and thus make the pixels easier to be identified. To evaluate the experimental results quantitatively, we multiply the factor of the pixels per degree of visual angle of the magnified headset screen with the measured length of the deviated position to obtain the reprojection error. Figure 8b presents an ablative analysis of hand optimization step. With updated hand models, our combined approach shows the best performance with the lowest average reprojection errors.

FIGURE 8 (a) PCK benchmark with the Stereo data set. The model trained with the improved data set (orange) shows a higher tracking accuracy compared with the original approach (blue). (b) Reprojection errors of occlusion masks acquired by different approaches for three sequences, pencil, pencil sharpener, and eraser. While using cost volume filtering (orange) to improve the raw depth (brown) shows better accuracy, our approach (gray and blue) shows a further improvement



While we outperform the previous approaches in the quantitative comparison (Figure 8b), we emphasize that our approach can also handle complex scenes that the hand cannot be labeled as either foreground or background object.

6.4 | User study

We designed a participant-based cooperative qualitative evaluation to evaluate our application for real-object enhancement. Nine subjects (ages 18–26 years, average 21.7 years) without VR/AR experience participated in this study. The main goal of this study is to test sense of presence, realism of experience, stability, and identify potential issues through interviews after each trial. This study compares experiences of using four different conditions (Figure 9): (a) without any occlusion handling; (b) with a naive approach[22] that adjusts the transparency of virtual objects based on the angle of the palm; (c) render occluded objects without the updated hand model using the proposed method, and (d) render occluded objects with the updated hand model.

During the experiment with the configuration shown in Figure 6, each user went through three scenes interacting with a pencil, a box, a card, with four different conditions. The sequence of trials in each scene was randomized to prevent bias. Users followed instructions to perform the simple task of interacting with objects with translating and rotating motions. After each trial, feedback was collected through a semistructured interview.

Figure 10 illustrates the results of the study. From both the results of the questionnaire as well as the comments in the subsequent open discussions, we confirmed a positive impression of our implementation. Since most of objects are partly occluded by fingers during interaction, and participants were actually holding realistic objects in hand, the naive solution of adjusting transparency to be fully opaque when participants flip their hands outward was highly problematic under this situation and resulted in a strange impression that virtual objects seemed to be fading away when they turned their arm. This problem can be observed in Figure 9. With a K–W test, we verified a more immersive MR experience and a

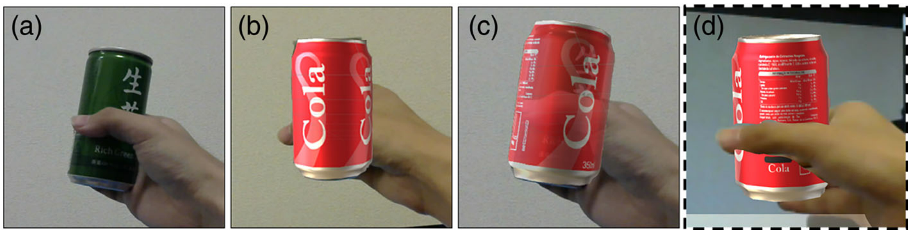


FIGURE 9 Results of three methods to augment a green can with a virtual Cola can: (a) the real scene without any overlay; (b) result without any occlusion handling; (c) result when applying the approach proposed by Liang et al.[22] The transparency was adjusted to 70% according to the direction of the palm in this case; (d) result of our method

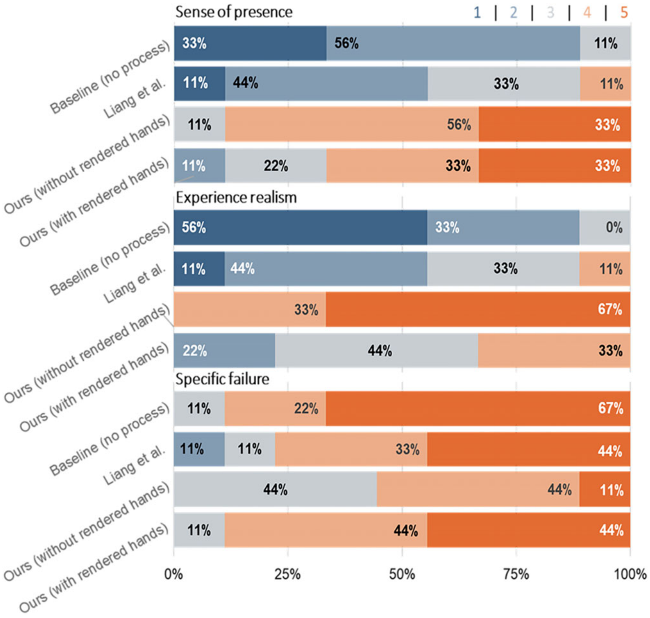


FIGURE 10 Likert-type survey result from the user study. The experience of each trial is rated from 1 as "*Bad*" to 5 as "*Good*" with a step of 1

significantly more realistic feeling of interacting with virtual objects with our approach ($p < .001$ in Scenes 1 and 3, $p = .022$ in Scene 2). Some users reported that rendering model hands mitigates some latency and resulted asynchronization, and thus give them a more consistent feeling.

7 | CONCLUSION AND DISCUSSION

In this article, we have presented a real-time method to handle the hand-object occlusions in MR. We propose a photorealistic RGB-D hand data set with precise joint and segmentation annotations to facilitate our occlusion-aware joint learning system. With a novel real-time optimization pipeline, we utilize the jointly estimated poses and masks to calculate occlusion masks and render objects with correct occlusions. The experimental results show better quantitative and qualitative performance than previous literature, and a user study verifies a more realistic MR experience of hand-object interactions. The implementation shows good accuracy, robustness, and speed with the potential to be further adapted to other applications.

Since we are using a commercial implementation of object augmentation this time, there is a technical issue of misalignment when localizing the optimized hand model. We believe a reimplementation can solve this problem. As a general limitation of learning-based approaches, greatly changing the appearance of hands such as wearing gloves may reduce the robustness. In addition, there is no sophisticated occlusion-aware object tracking in the current implementation, and this lead to losing augmentation of the object during experiments due to strong occlusions. Joint tracking of hand and object is a promising direction for future improvements.

ACKNOWLEDGEMENT

This research was supported by the JST ACCEL (JPMJAC1602), JST-Mirai Program (JPMJMI19B2), JSPS KAKENHI (JP17H06101, JP19H01129), and the Royal Society (IES\R2\181024).

ORCID

Qi Feng https://orcid.org/0000-0002-6892-3122

Hubert P. H. Shum https://orcid.org/0000-0001-5651-6039

REFERENCES

1. Zhang E, Cohen MF, Curless B. Emptying, refurnishing, and relighting indoor spaces. ACM Trans Graph. 2016;35(6):1–14.
2. Gupta A, Cecil J, Pirela-Cruz M, Ramanathan P. A virtual reality enhanced cyber-human framework for orthopedic surgical training. IEEE Syst J. 2019;13(3):3501–3512.
3. Holynski A, Kopf J. Fast depth densification for occlusion-aware augmented reality. ACM Transactions on Graphics. 2019;37(6):1–11.
4. Tian Y, Guan T, Wang C. Real-time occlusion handling in augmented reality based on an object tracking approach. Sensors. 2010;10(4):2885–2900.
5. Chao D Chen YL, Ye M, Ren L. Edge snapping-based depth enhancement for dynamic occlusion handling in augmented reality. Proceedings of the 2016 IEEE International Symposium Mixed and Augmented Reality, image. Merida, Mexico; 2016. p. 54–62.
6. Azuma RT. Making augmented reality a reality. Imaging and applied optics 2017. Rochester, New York: Optical Society of America, 2017; JTu1F.1.
7. Lepetit V, Berger MO. Handling occlusion in augmented reality systems: A semi-automatic method. Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000), Munich, Germany; 2000. p. 137–146.
8. Tian Y, Long Y, Xia D, Yao H, Zhang J. Handling occlusions in augmented reality based on 3d reconstruction method. Neurocomputing. 2015;156:96–104.
9. Patney A, Kim J, Salvi M, et al. Perceptually-based foveated virtual reality. ACM SIGGRAPH 2016 emerging tech. New York, NY, United States: Association for Computing Machinery, 2016; p. 17.
10. Walton DR Steed A. Accurate real-time occlusion for mixed reality. Proceedings of the 23rd ACM Symposium Virtual Reality Software and Technology, Gothenburg, Sweden; 2017. p. 11.
11. Kyriazis N, Oikonomidis I, Panteleris P, et al. A generative approach to tracking hands and their interaction with objects. Man–Machine Interactions. Volume 4 Cham: Springer; , 2016; p. 19–28.
12. Tagliasacchi A, Schröder M, Tkach A, Bouaziz S, Botsch M, Pauly M. Robust articulated-icp for real-time hand tracking. Comp Graph Forum. 2015;34:101–114.
13. Qian C, Sun X, Wei Y, Tang X, Sun J. Realtime and robust hand tracking from depth. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ohio, United States; 2014. p. 1106–1113.
14. Mueller F, Mehta D, Sotnychenko O, Sridhar S, Casas D, Theobalt C. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy; 2017.

15. Romero J, Tzionas D, Black MJ. Embodied hands: Modeling and capturing hands and bodies together. ACM Trans Graph. 2017;36(6):245:1–245:17.
16. Mueller F, Davis M, Bernard F, et al. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. ACM Trans Graph. 2019;38(4):49.
17. Tekin B Bogo F, Pollefeys M. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, California, United States; 2019. p. 4511–4520.
18. Simon T Joo H, Matthews I, Sheikh Y. Hand keypoint detection in single images using multiview bootstrapping. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, United States; 2017. p. 1145–1153.
19. Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C. Ganerated hands for real-time 3d hand tracking from monocular rgb. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Utah, United States; 2018. p. 49–59.
20. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognitiones. Florida, United States: IEEE; 2009. p. 248–255.
21. Zhang J, Jiao J, Chen M, Qu L, Xu X, Yang Q. 3d hand pose tracking and estimation using stereo matching; 2016. arXiv preprint arXiv:1610.07214.
22. Liang H, Yuan J, Thalmann D, Thalmann NM. Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia; 2015. p. 743–744.
23. Lu Y, Smith S. GPU-Based Real-Time Occlusion in an Immersive Augmented Reality Environment. Journal of Computing and Information Science in Engineering. 2009;9: 2.
24. Dong S, Kamat V. Resolving incorrect visual occlusion in outdoor augmented reality using TOF camera and OpenGL frame buffer. Proceedings of International Conference on Construction Applications of Virtual Reality. 2010;55–64.

AUTHOR BIOGRAPHIES

Qi Feng is currently pursuing his Ph.D. degree in Waseda University. He received the B.E. and M.E. in Applied Physics from the Graduate School of Advanced Science and Engineering at Waseda University, Tokyo, Japan in 2017 and 2019, respectively. His main research area includes deep learning applications, computer vision, computer graphics, virtual and augmented reality.

Hubert P.H. Shum is an associate professor (reader) in Computer Science at Northumbria University and the Director of Research and Innovation of the Computer and Information Sciences Department. Before that, he was a Senior Lecturer at Northumbria University, a Lecturer at the University of Worcester and a postdoctoral researcher at RIKEN Japan. He received his Ph.D. degree from the University of Edinburgh, his Master and Bachelor degrees from the City University of Hong Kong.

Shigeo Morishima is a professor of Graduate School of Advanced Science and Engineering, Waseda University in Japan. He received the B.S., M.S., and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, Japan, in 1982, 1984, and 1987, respectively. From 1987 to 2001, he was an associate professor and from 2001 to 2004, a professor of Seikei University. He is a member of the IEEE, ACM SIGGRAPH, and is a trustee of Japanese Academy of Facial Studies. He is a vice president of the Institute of Image Electronics Engineers of Japan.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Feng Q, Shum HPH, Morishima S. Resolving hand-object occlusion for mixed reality with joint deep learning and model optimization. *Comput Anim Virtual Worlds*. 2020;e1956. https://doi.org/10.1002/cav.1956