Mechatronics 110 (2025) 103365

Contents lists available at ScienceDirect

Mechatronics

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

Testing human-hand segmentation on in-distribution and out-of-distribution data in human–robot interactions using a deep ensemble model

Reza Jalayer [a,b,*], Yuxin Chen [b], Masoud Jalayer [c], Carlotta Orsenigo [a], Masayoshi Tomizuka [b]

[a] Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 24/b, 20156, Milan, Italy
[b] Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA 94709, USA
[c] Department of Materials and Mechanical Engineering, University of Turku, Vesilinnantie 5, Turku, 20014, Finland

ARTICLE INFO

Keywords:
Human–robot interaction
Uncertainty in AI
Hand segmentation
Industry 5.0
Hand gesture
Ensemble models

ABSTRACT

Reliable detection and segmentation of human hands are critical for enhancing safety and facilitating advanced interactions in human–robot collaboration. Current research predominantly evaluates hand segmentation under in-distribution (ID) data, which reflects the training data of deep learning (DL) models. However, this approach fails to address out-of-distribution (OOD) scenarios that often arise in real-world human–robot interactions. In this work, we make three key contributions: first we assess the generalization of deep learning (DL) models for hand segmentation under both ID and OOD scenarios, utilizing a newly collected industrial dataset that captures a wide range of real-world conditions including simple and cluttered backgrounds with industrial tools, varying numbers of hands (0 to 4), gloves, rare gestures, and motion blur. Our second contribution is considering both egocentric and static viewpoints. We evaluated the models trained on four datasets, i.e. EgoHands, Ego2Hands (egocentric mobile camera), HADR, and HAGS (static fixed viewpoint) by testing them with both egocentric (head-mounted) and static cameras, enabling robustness evaluation from multiple points of view. Our third contribution is introducing an uncertainty analysis pipeline based on the predictive entropy of predicted hand pixels. This procedure enables flagging unreliable segmentation outputs by applying thresholds established in the validation phase. This enables automatic identification and filtering of untrustworthy predictions, significantly improving segmentation reliability in OOD scenarios. For segmentation, we used a deep ensemble model composed of UNet and RefineNet as base learners. Our experiments demonstrate that models trained on industrial datasets (HADR, HAGS) outperform those trained on non-industrial datasets, both in segmentation accuracy and in their ability to flag unreliable outputs via uncertainty estimation. These findings underscore the necessity of domain-specific training data and show that our uncertainty analysis pipeline can provide a practical safety layer for real-world deployment.

1. Introduction

In the era of Industry 4.0, industrial robots and humans increasingly work side-by-side in collaborative environments, moving away from the traditional approach of isolating robots within cages to prevent accidents [1,2]. This shift underscores the critical importance of human safety, particularly in preventing or mitigating collisions in shared workspaces [3]. Hands are among the most vulnerable parts of the human body in these interactions which are constantly present in the proximity of industrial robots [4]. Reliable detection and segmentation of human hands are therefore essential to ensure human safety as a main objective. Beyond safety — one of the cornerstones of Industry 4.0 — Industry 5.0 expands the focus to emphasize human-centric interactions. This newer approach encourages manufacturers to prioritize not only safety but also the comfort and well-being of operators during human–robot interactions [5]. In this context, enabling robots to understand human hand gestures and actions through accurate hand segmentation has become an area of growing research interest [6].

There are various approaches to detecting human hands around industrial robots, ranging from wearable sensors to computer vision-based methods. Wearable sensors, such as gloves equipped with tracking devices [7,8] or markers for positional detection [9], have been explored in previous studies. However, the constant need to wear

* Corresponding author at: Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 24/b, 20156, Milan, Italy.
E-mail address: reza.jalayer@polimi.it (R. Jalayer).

https://doi.org/10.1016/j.mechatronics.2025.103365
Received 30 December 2024; Received in revised form 1 May 2025; Accepted 1 June 2025
Available online 21 June 2025

these devices during operation can be frustrating for workers, and the setup requirements before use further limit their practicality in industrial environments. In contrast, computer vision-based techniques offer a more operator-friendly solution, allowing seamless hand detection and segmentation without requiring workers to wear any additional equipment. The rapid advances in Machine Learning (ML) and, more specifically, Deep Learning (DL), coupled with the growing availability of powerful hardware, have further propelled computer vision as the preferred choice for hand detection and segmentation. Unlike traditional computer vision methods, such as template matching [10], which demand manual intervention and yield less accurate results, DL-based models deliver superior accuracy and robustness, making them more suitable for complex industrial applications [6].

DL-based hand segmentation models heavily depend on the datasets they are trained on [11]. While several hand segmentation datasets have been introduced in the literature and made publicly accessible, significant challenges arise when applying these models to human–robot interactions in industrial domains. One primary challenge is that the backgrounds in these datasets are often biased towards accessible environments, such as residential settings or scenes containing common household objects [12]. This bias makes it difficult for models trained on such datasets to generalize to industrial scenarios with complex and cluttered backgrounds. Another challenge lies in the limited number of hands present in most datasets, typically restricted to one or two per scene. However, in real-world interactions with robots, multiple participants or hands may simultaneously be in the frame. Furthermore, these datasets often capture hand images using either egocentric and mobile or static cameras, while real-world scenarios may involve different camera setups that deviate from these idealized conditions. Additionally, the gestures captured in these datasets are often casual and fall within predefined categories, overlooking the diversity and complexity of hand gestures in actual industrial interactions. For instance, gestures such as finger crossings or interactions with robotic hands are rarely represented. Moreover, most datasets predominantly feature bare human hands, ignoring the fact that industrial operators typically wear gloves for safety, which alters the appearance and texture of hands [12].

On the other hand, creating new datasets that address these challenges is a labor-intensive process, requiring significant time and resources. Even if such datasets were created, their generalization to other human–robot tasks and conditions would remain uncertain. Consequently, there is a notable gap in the literature exploring how well pre-trained models, trained on existing datasets, perform when applied to real-world industrial hand segmentation tasks.

In this study, we address these challenges by conducting a novel investigation into the performance of DL models for hand segmentation under both in-distribution (ID) and out-of-distribution (OOD) conditions. We train models using non-industrial datasets, such as EgoHands [13] and Ego2Hands [14], as well as an industrial-like dataset, HAGS and HADR [12,15]. To evaluate these models, we collect real-world images of human–robot interactions in industrial environments, encompassing a wide range of scenarios. These include ID conditions that align with the training datasets and OOD conditions that introduce novel challenges, such as complex gestures.

To ensure accurate evaluations, we use a deep ensemble model, an uncertainty-aware DL approach [16]. By ensembling segmentation models that have been successfully used in the field, we create a deep ensemble model to quantify model uncertainty on both ID and OOD data as well as assessing the accuracy of hand segmentation.

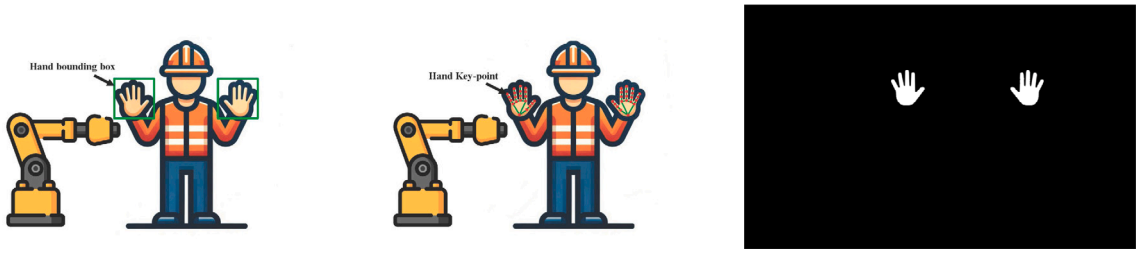
The remainder of this paper is structured as follows: In Section 2, we provide a detailed overview of related works, highlighting previous approaches, datasets, and the challenges of applying existing methods to industrial human–robot interactions. Section 3 describes the datasets and experimental setup, detailing the training datasets, test data collection, and experimental methodology, including details of segmentation models and metrics. In Section 4, we present the results of our experiments, analyzing model performance on different data conditions, insights from segmentation performance and uncertainty quantification as well as qualitative results. Finally, in Section 5, we summarize our findings and propose directions for future research.

2. Related works

In vision-based hand recognition, existing studies can generally be categorized based on their primary focus, as schematically shown in Fig. 1, into three groups: hand detection (Fig. 1(a)), hand key-points detection (Fig. 1(b)), and hand segmentation (Fig. 1(c)). Hand detection typically involves locating a bounding box around the region containing hands, commonly achieved by employing deep learning (DL) architectures such as models from the YOLO (You Only Look Once) family [17]. This approach has been extensively used in human–robot interaction (HRI) applications for identifying the presence and approximate position of human hands within a workspace [18,19]. Hand key-point detection aims to identify specific landmarks or critical points on the hand, typically using established, open-source libraries such as OpenPose [20] or MediaPipe [21]. Such key-point-based methods have been effectively leveraged in HRI studies to recognize hands facilitate interactions between humans and robots [22,23]. However, hand segmentation (the primary focus of our study) involves a pixel-level identification of the hand region, providing a more granular and precise separation of hands from their surrounding environment. This precise segmentation is crucial for improving the accuracy of gesture recognition (in a possible next step after hand segmentation) and enhancing safety in industrial human–robot collaborative scenarios.

Hand segmentation have gained significant attention in recent years, particularly in computer vision-based applications. These techniques have been applied to a variety of domains, including medical purposes such as rehabilitation tasks [24] and sign language detection for speech-impaired individuals [25,26]. They are also widely used in human–computer interaction scenarios, especially with the advancement of augmented reality (AR), virtual reality (VR), and mixed reality (MR) technologies [27,28]. Within the context of human–robot interaction (HRI), hand segmentation based on computer vision techniques has facilitated applications such as gesture recognition for controlling mobile robots [29–31] and assisting surgical robots [32]. Despite this progress, the application of hand segmentation in industrial HRI has received comparatively less attention, where unique challenges demand tailored approaches.

While many recent works employ DL models for hand segmentation, these models face limitations when applied to industrial settings. There are few and recent works in human–robot interactions using hand segmentations as listed in Table 1. For instance, Sajedi et al. [33] leveraged a Bayesian Neural Network for hand segmentation, focusing on RGB images captured from a third-person perspective who carries a mobile phone to record the data in a human–robot interaction. Their model, trained on the EgoHands dataset [13], incorporated uncertainty quantification to improve segmentation accuracy. The hand instances in their data were restricted to two hands of the operator working with the robot. Also, the experimental context could be closer to reality by the presence of gloves or cluttered workspaces. Vysocky et al. [34], also generated their own data to segment human hands by UNet model in interaction with an industrial collaborative robot. Their test was restricted to only segment hands of one operator in the scene, with no gloves and no motion noise, while hands represent only casual gestures. Grushko et al. [15] introduced HADR, a synthetic RGB-D dataset designed for industrial applications, using domain randomization to mimic real-world conditions. Despite its contributions, HADR is limited by its static top-down camera angles and a restriction to two hand instances per frame, which fails to capture the dynamic and multi-user nature of industrial collaboration. Sharma et al. [12] recently addressed some of these gaps by introducing the HAGS dataset, which includes gloved and ungloved hands captured from stationary side and top-down cameras. However, the dataset also limits hand instances to two per frame and does not explore rare gestures or the influence of background complexity. Their results with segmentations with DL-based models e.g. UNet trained on the available non-industrial datasets resulted in



(a) Hand detection by bounding box (b) Hand keypoints detection (c) Hand segmentation

Fig. 1. Schematic visualizations of (a) hand detection by bounding box (b) hand key-points detection (c) hand segmentation in a human robotic interaction scene.

Table 1
Summary of related works, implemented hand segmentation in industrial HRI.

Paper	Year	DL model	Implementation	Limitation
[33]	2022	Bayesian Neural Networks (BNN)	Implementing hand segmentation in diverse disassembly task with a UR5e Co-robot arm	- Limited to a maximum of 2 hands per frame. - No gloves, no motion blur, only casual gestures. - Unspecified camera angles (recorded via mobile phone).
[34]	2022	UNet	Testing hand segmentation to segment hands interacting with an industrial collaborative robot (UR3e)	- Limited to a maximum of 2 hands per frame. - No gloves, no motion blur, only casual gestures. - Only one viewpoint (top-down and static).
[15]	2023	MaskRCNN	Building a synthetic hand segmentation dataset in industrial backgrounds with the presence of a robotic arm	- Synthetic data only, limited to 2 hands per frame. - No gloves, no motion blur, only casual gestures. - Only one viewpoint (top-down and static).
[12]	2024	UNet	Building a real hand segmentation dataset across assembly tasks with UR3e industrial arm in glovebox environments	- Limited to a maximum of 2 hands per frame. - Not considering diverse industrial backgrounds. - No motion blur, only casual gestures. - Cameras are static

non-satisfactory results emphasizing the importance of creating more domain-specific datasets like HAGS for future research.

Based on these works, several critical limitations and challenges emerge. First, existing datasets for hand segmentation in HRI restrict the number of hands to two, overlooking multi-user industrial scenarios where more than two hands may be present. Second, the gestures captured in these datasets are often casual and lack diversity, failing to account for rare but practical gestures such as interlocked fingers or crossed hands. Third, no studies considered noisy conditions such as motion blur, which often occur during dynamic interactions with robots. Additionally, while gloves are commonly used in industrial settings, only HAGS, includes gloved hands. Finally, existing studies predominantly rely on static cameras, with no research on considering both egocentric and static cameras to provide a better evaluation.

Despite these limitations, creating new datasets tailored to industrial HRI remains a labor-intensive task, and questions about their generalization to diverse scenarios persist. There is also a lack of studies evaluating the performance of pre-trained models in real-world industrial conditions. Moreover, the relationship between segmentation accuracy and background complexity (e.g., cluttered environments vs. object-free scenes) has been largely unexplored in the literature, which limits our understanding of the robustness of existing models.

Our study addresses these gaps by testing the performance of state-of-the-art DL models on a test dataset that incorporates both in-distribution (ID) and out-of-distribution (OOD) conditions. This work examines challenging scenarios, including rare gestures, motion blur, gloved and non-gloved hands across a combination of egocentric and static camera perspectives. By leveraging uncertainty-aware deep ensemble models, we aim to provide an evaluation of hand segmentation performance and contribute insights for future advancements in industrial human–robot interactions.

3. Datasets and experimental setup

This section provides details on the datasets used for training and testing the deep learning models, as well as the experimental setup designed to evaluate their performance. We describe the datasets we choose to train our models, our custom-designed test dataset for in-distribution (ID) and out-of-distribution (OOD) evaluation, and the methodology and pipeline of our work.

3.1. Training datasets

For training the deep learning (DL) models, we considered both industrial and non-industrial datasets. This was done because on one hand there is no hand segmentation dataset with egocentric view in industrial-like context and on the other considering trained model on industrial and non-industrial images gives us a better evaluation for our results. The summary of the description of each training dataset is provided in Table 2.

Since no industrial-like datasets exist for hand segmentation in egocentric views, we utilized two widely recognized non-industrial datasets: EgoHands and Ego2Hands. EgoHands is a very large available egocentric dataset for hand segmentation, comprises 4800 pixel-level annotated frames captured using Google Glass. The frames depict diverse non-industrial backgrounds and contain between zero to four hands per image. This dataset has been widely employed for training hand segmentation models in a variety of applications, including human–robot interaction. However, it lacks relevance to industrial contexts, such as the use of gloves or cluttered backgrounds, which limits its applicability to more complex environments. Ego2Hands is a more recent egocentric dataset with 2000 annotated frames of RGB images, each containing a single person and up to two hand instances. This dataset employs background replacement techniques, providing greater control over environmental diversity compared to EgoHands. Additionally, it introduces inter-occluded hands for the first time and ensures a more even spatial distribution of hand positions across frames, addressing some limitations of its predecessor. Despite these advancements, Ego2Hands remains focused on non-industrial settings, with backgrounds and hand interactions that do not reflect the complexities of industrial human–robot interaction scenarios.

To incorporate industrial perspectives, we included two datasets designed specifically for such contexts: HADR and HAGS. HADR is a

Table 2
Summary of four training datasets.

Training dataset	Point of view	Industrial	Real/Synthetic	RGB/RGB-D	Max No. of hands per frame	No. of annotated frames
EgoHands [13]	Egocentric	No	Real	RGB	4	4800
Ego2Hands [14]	Egocentric	No	Real	RGB	2	2000
HADR [15]	Top-down	Yes	Synthetic	RGB-D	2	117 000
HAGS [12]	Top-Down + Side	Yes	Real	RGB	2	1728

synthetic dataset with RGB-D annotations, created using domain randomization techniques to reduce the reality gap between synthetic and real-world data. This was achieved by introducing random variations in distractor object properties, background textures, camera positions and orientations, and lighting conditions. The dataset contains 117,000 annotated frames, making it one of the largest datasets available for hand segmentation. For our study, we used only the RGB modality of HADR, as our test data, described later, was captured using RGB cameras. Each frame in HADR contains a maximum of two hand instances, all viewed from a top-down perspective. While the dataset is comprehensive in terms of synthetic diversity, its static camera viewpoint and synthetic nature limit its application in fully dynamic or real-world settings. HAGS, on the other hand, provides a realistic dataset consisting of 1728 frames with pixel-level annotations of human hands interacting with industrial robots. The dataset includes both gloved and ungloved hands, addressing a key limitation of many previous datasets. Images in HAGS were captured using two stationary cameras: a GoPro Hero 7 capturing top-down views and a RealSense Development Kit Camera SR300 capturing side views. To increase variability, the dataset incorporated background replacement techniques, using both real and synthetic backgrounds. However, like HADR, HAGS restricts the number of hands per frame to two instances and focuses on industrial scenarios while overlooking the presence of some challenging conditions like rare hand gestures.

To provide further insights into the datasets, we report the relative frequency distribution of the hand instance sizes (defined as the ratio of hand pixels to total image pixels) for each dataset in Fig. 2. As evident in this figure, for all four training datasets, most of hand instances occupy less than 0.1 of the total image pixels. Additionally, the hand instances in the HAGS, EgoHands, and HADR datasets generally occupy smaller image areas compared to those in Ego2Hands. This difference can be attributed to both camera placement and experimental conditions. Specifically, HAGS and HADR datasets are captured from static viewpoints from a distance to encompass the entire human–robot interaction scene. Consequently, human hands do not occupy larger portions of the images, limiting the presence of large hand instances. Also, the EgoHands dataset involves an egocentric, mobile viewpoint of a participant performing joint activities with another person. This collaborative experimental setup naturally restricts hand instances from becoming excessively large within each frame. In contrast, Ego2Hands captures participants performing free-hand motions directly in front of a head-mounted egocentric webcam, allowing the dataset to encompass a wider range of hand instance sizes.

3.2. Test dataset design

To evaluate the trained models, we created a realistic image dataset by recording interactions with an industrial robotic arm (FANUC LR Mate 200iD 7L robot). The dataset was captured from two camera perspectives: a static side-view camera and an egocentric camera mounted on the operator's head. This dual-camera setup was chosen to provide comprehensive coverage of the interaction scene and to evaluate model performance across varying viewpoints.

For the side view, we used an Intel RealSense D435 camera, positioned on the right-hand side of the robot, as shown in the left of Fig. 3. The camera angle was carefully adjusted to capture the interaction space, including the operator's hands and the robotic arm, ensuring that the entire workspace was visible. For the egocentric view, a GoPro camera was mounted on the operator's helmet using a headband, allowing us to capture the perspective of the operator during the interaction as evident in the left picture of Fig. 3.

The dataset includes diverse interaction scenarios with one and two operators closely working with the robot. To ensure varied background complexity, we recorded videos in both object-free environments and cluttered industrial settings with industrial tools (e.g. hammer, scissors, wrenches, nuts and bolts, ...) present. Hands were captured with and without gloves to reflect realistic industrial conditions, where operators frequently wear gloves for safety. Additionally, some videos intentionally included rare hand gestures (e.g., interlocked fingers and crossed hands) to create out-of-distribution (OOD) data for testing in these conditions. Motion-blurred frames caused by fast-moving hands were also included to simulate real-world challenges as aleatoric uncertainty.

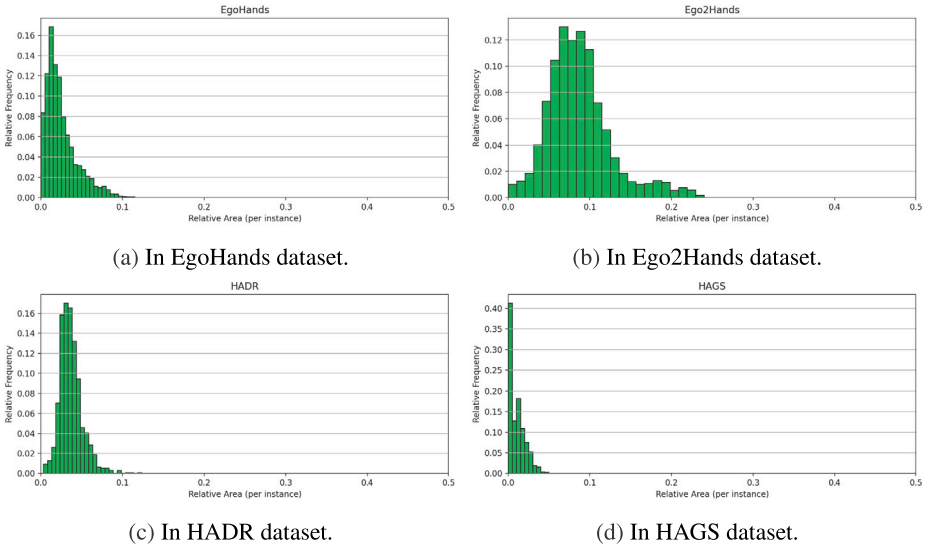
To streamline the description of dataset conditions, we use the following abbreviations: one operator (O1), two operators (O2), gloved hands (GH), hands with rare gestures (RG), and motion-blurred noisy hands (MBN). Fig. 4 illustrates these conditions for clarity.

3.3. In-distribution (ID) and out-of-distribution (OOD) scenarios

For effective evaluation of deep learning (DL) models, it is crucial to assess their performance not only on data that aligns with their training distribution (in-distribution (ID) data) but also on data that deviates from this distribution (out-of-distribution (OOD) data). The discrepancy between ID and OOD data introduces uncertainty in the model's predictions, which can significantly affect their reliability. Understanding and addressing this uncertainty is essential, especially in safety-critical applications like human–robot interaction. Uncertainty in predictions can arise from two main sources: epistemic uncertainty and aleatoric uncertainty [35]. Epistemic uncertainty, often referred to as knowledge uncertainty, occurs when the model has not encountered all the characteristics of the data during training. This type of uncertainty can be reduced by introducing additional training data that better represents these unseen scenarios. On the other hand, aleatoric uncertainty, also known as data uncertainty, arises due to inherent variability in the data, such as noise or measurement errors. Unlike epistemic uncertainty, aleatoric uncertainty is irreducible and cannot be mitigated by adding more training data [16].

In our experiments, OOD data stemming from epistemic uncertainty includes conditions absent in the training datasets, such as gloved hands (GH), rare gestures (RG), and the presence of two human operators (O2), when the training data contained only one operator. Aleatoric uncertainty arises from noisy conditions, such as motion-blurred hands (MBN), which were not included in the training datasets. These conditions were specifically designed to evaluate the robustness and generalization capabilities of the models, as summarized in Table 3.

As shown in Table 3, certain scenarios are consistently classified as OOD across all training datasets. For instance, rare gestures (RG) are absent in all datasets, making them OOD for all trained models. Similarly, motion-blurred hands (MBN), representing aleatoric uncertainty, are not included in any training datasets and are treated as OOD for all models. However, some scenarios depend on the specific training dataset. For example, gloved hands (GH) are considered ID for models trained on HAGS, as this dataset includes gloved hands during training, but OOD for models trained on EgoHands, Ego2Hands, and HADR. Likewise, two human operators (O2) are considered ID for models trained on EgoHands, as this dataset includes scenes with multiple operators, but OOD for other datasets where the presence of two operators is absent.



(a) In EgoHands dataset. (b) In Ego2Hands dataset.

(c) In HADR dataset. (d) In HAGS dataset.

Fig. 2. Distribution of relative hand instance sizes for the training datasets.

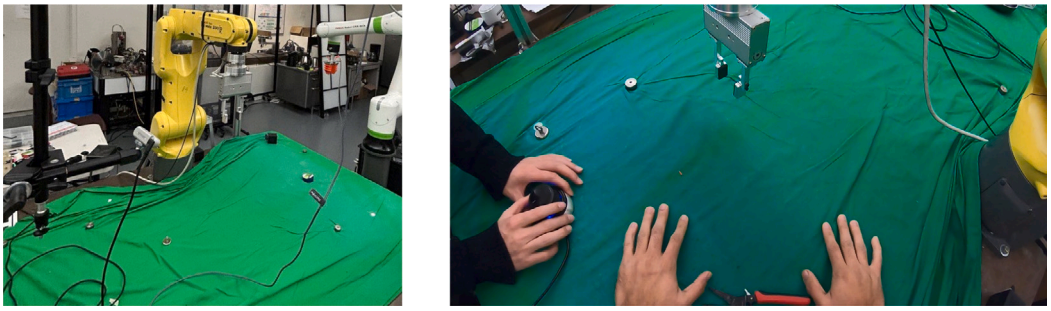


Fig. 3. Dataset collection setup. We mount an Intel RealSense D435 on the right-hand side of the robotic arm for static images (left), and a GoPro camera on the helmet of the operator to capture egocentric images (right).

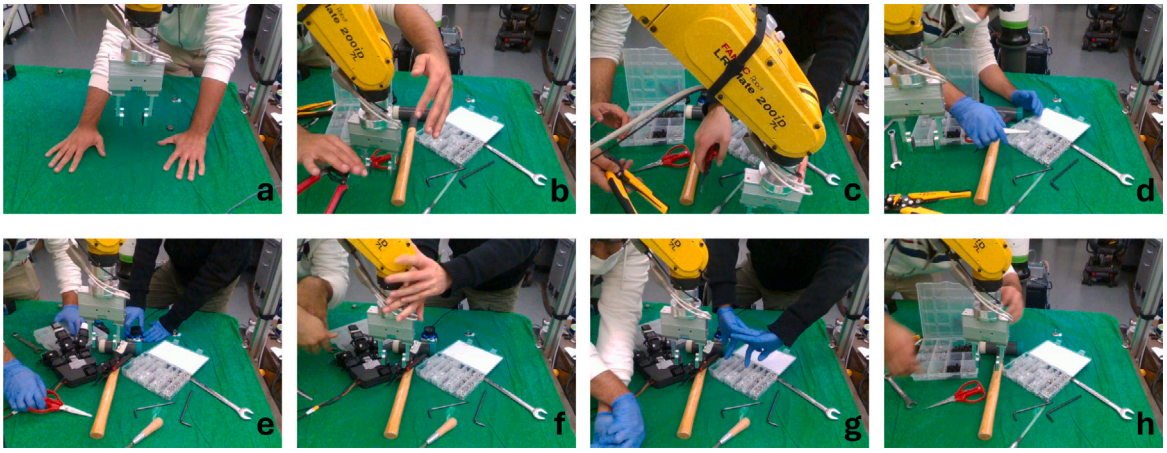


Fig. 4. Different scenarios of human–robot interactions from the side view static camera: one operator (O1) in (a) simple background and (b) cluttered background, (c) presence of two operators (O2), (d) one operator (O1) with gloves (GH), (e) two operators (O2) with gloves (GH), (f) two operators (O2) with rare gestures (RG), (g) two operators (O2) with gloves (GH) having rare gestures (RG), and (h) image with motion blurred noise (MBN).

3.4. Data preparation

To construct the test dataset, we recorded nine videos capturing interactions between one and two operators with an industrial robotic arm, using both side-view and egocentric cameras. In total, 34,577 frames were recorded from each camera and each frame for the side camera has the dimensions of 640 × 480 pixels and for the egocentric camera 1920 × 1080 pixels. Since the videos were captured using high-frame-rate cameras, many frames were repetitive, with minimal or no change between consecutive frames, making them redundant for our analysis. To address this, we conducted a manual review of the recorded videos, carefully selecting frames that were non-repetitive and meaningful for our study. After this refinement process, 1871 unique frames were retained from each camera (side-view and egocentric).

Table 3
OOD scenarios of the training dataset.

Training dataset	OOD scenarios	
	Epistemic uncertainty	Aleatoric uncertainty
EgoHands [13]	Gloved hands (GH), rare gestures (RG)	Motion-blurred noisy hands (MBN)
Ego2Hands [14]	Two human operators (O2), gloved hands (GH), rare gestures (RG)	Motion-blurred noisy hands (MBN)
HADR [15]	Two human operators (O2), gloved hands (GH), rare gestures (RG)	Motion-blurred noisy hands (MBN)
HAGS [12]	Two human operators (O2), rare gestures (RG)	Motion-blurred noisy hands (MBN)

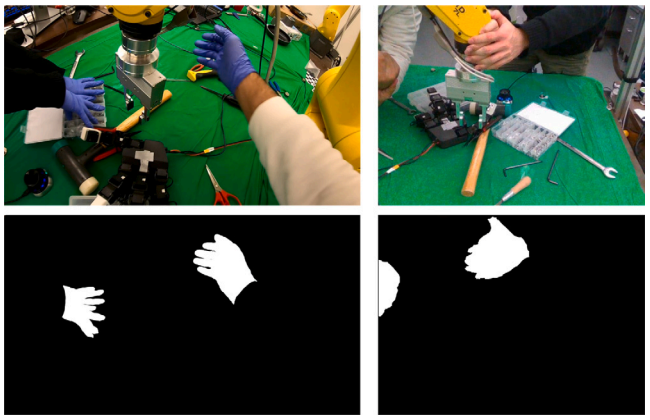


Fig. 5. Sample frames with ground truth masks, captured from egocentric view (left) and side view camera (right).

For the testing phase, we aimed to achieve a balanced dataset across the 8 conditions defined in Fig. 4. From each condition, 20 frames were randomly selected for annotation, resulting in a total of 320 annotated images (160 frames each for the side and egocentric views).

The selected images were annotated with pixel-wise accuracy for binary classification, where hand regions were labeled as the foreground and all other areas as the background. To achieve high-quality annotations, we utilized Label Studio [36], an open-source data labeling tool, to create initial coarse annotations. These annotations were meticulously refined manually to ensure precision and consistency across the dataset. As the trained models are designed to segment only the hand region — from the tips of the fingers to the wrist — we adhered to this definition, excluding the arm and forearm from the annotated hand regions to maintain alignment with the training datasets. Fig. 5 illustrates examples of annotated images, showcasing the accuracy and uniformity of the labeling process.

This rigorous data preparation workflow resulted in a dataset that is not only diverse and representative of various conditions but also meticulously balanced across all categories. By ensuring high annotation quality and meaningful representation of scenarios, the dataset is well-suited for comprehensively evaluating the performance of DL models under different conditions.

For additional information about the test data, we present the relative frequency distribution of hand instance sizes for both egocentric and side-view images in Fig. 6. As shown in this figure, similar to the training datasets (Fig. 2), most hand instances in both views of the test images occupy less than 0.1 of the total pixels. However, the egocentric images display a wider range of hand instance sizes, which can be attributed to the fact that hands in this viewpoint can appear closer to the camera, compared to the fixed side-view perspective.

3.5. Models and metrics

In this work, we used deep ensemble for the evaluations. Deep ensembles are consist of multiple of DL models called base learners which are trained independently to improve predictive accuracy and make the uncertainty of predictions quantifiable [16]. For the choice of base learners, we selected two widely recognized DL algorithms for segmentation: UNet [37] and RefineNet [38]. These models leveraged the encoder–decoder technique and have been successfully employed in prior works on human hand segmentation [12,14,39,40]. To justify the choice of these models over other existing segmentation models, UNet has been shown to have better segmentation accuracy results over other recent segmentation models, e.g., MobileSAM [41] and BiSeNetv2 [42] in a HRI study on the HAGS dataset [12]. It was also considerably faster due to its lightweight architecture (more than 6 times faster than MobileSAM and 3 times faster than BiSeNetv2), reinforcing the choice of UNet. Also, RefineNet, was used as a selected segmentation model in the Ego2Hands study [14], showing its high generalization accuracy for cross-dataset evaluation. Additionally, RefineNet outperformed other baseline segmentation models in the study using EgoHands in segmenting hands [40].

The configuration of this model is illustrated in Fig. 7. in this configuration, DE-Mix, combines both UNet and RefineNet models as base learners. This approach incorporates the concepts of the heterogeneous ensembles (different base learners), as discussed in prior ensemble learning research [43]. Using heterogeneous ensembles, have been shown to enhance performance by leveraging the diversity of base learners [44]. While deep ensemble models have been applied to hand segmentation tasks before [12], to the best of our knowledge, the prior work has employed only homogeneous ensembles combining identical DL architectures for hand segmentation, such as ensemble of UNet [12].

As shown in Fig. 7, the output of the deep ensemble is derived from the predictions of K-base learners. Since we do binary classification of segmenting hands from background, each base learner performs binary pixel-wise classification on the input image, determining whether each pixel belongs to the background or a hand. The final output of the deep ensemble is computed as the average prediction across the K-base learners.

To evaluate segmentation performance, we use the mean Intersection over Union (mIoU), a widely accepted metric that quantifies the overlap between predicted and ground truth hand regions. This metric provides a comprehensive measure of segmentation accuracy by comparing the predicted segmentation mask to the ground truth.

To quantify the uncertainty of predictions, we use the average predictive entropy, a standard metric that evaluates the uncertainty associated with each pixel's prediction across all classes. For a given pixel p, the predictive entropy $E(p)$ is defined as:

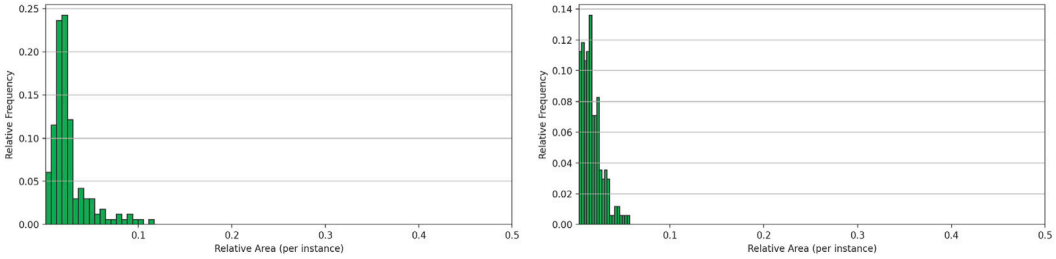
$$E(p) = -\sum_{c=1}^{C} P(c|p) \log P(c|p) \tag{1}$$

where $P(c|p)$ is the predicted probability of class c for pixel p. Since this is a binary segmentation task (hand and background), the number of classes is $C = 2$.

For deep ensemble models, the predicted probability $P(c|p)$ is computed as the average of the output scores from K base learners:

$$P(c|p) = \frac{1}{K} \sum_{k=1}^{K} P_k(c|p) \tag{2}$$

where $P_k(c|p)$ is the predicted probability from the kth base learner. The predictive entropy provides a pixel-level measure of uncertainty, with higher values indicating greater uncertainty in the model's prediction



(a) In egocentric view. (b) In side view.

Fig. 6. Distribution of relative hand instance sizes for the test datasets.

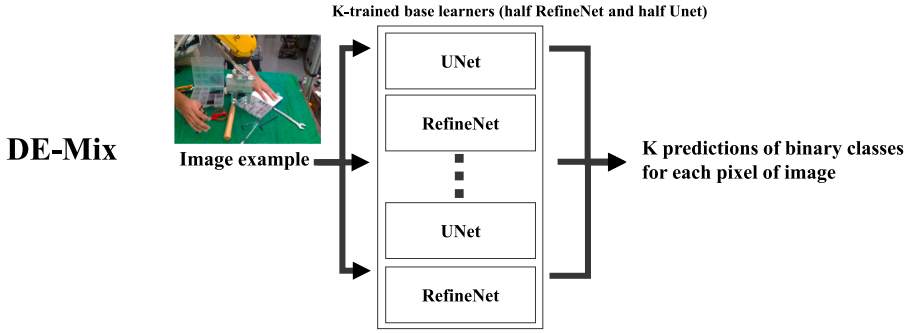


Fig. 7. Deep ensemble architectures (DE-Mix) made by K-trained models which half are UNet and RefineNet.

for that pixel. The average predictive entropy for the entire image, denoted as $\bar{E}$, is calculated as:

$$\bar{E} = \frac{1}{N}\sum_{p=1}^{N} E(p) \tag{3}$$

where N is the total number of pixels in the image.

To further evaluate uncertainty specific to the hand regions, we calculate the average predictive entropy for ground truth hand pixels, denoted as $\bar{E}_h$, which focuses only on pixels that belong to the ground truth hand regions. This measure is computed as:

$$\bar{E}_h = \frac{1}{N_h}\sum_{p\in H} E(p) \tag{4}$$

where N_h is the total number of ground truth hand pixels, and H represents the set of pixels corresponding to the hand regions in the ground truth. This metric helps evaluate whether the model exhibited uncertainty specifically in predicting the hand regions.

The previously defined metric, average predictive entropy of ground truth hand pixels ($\bar{E}_h$), is suitable for post-prediction analyses where ground truth labels are available. However, for uncertainty quantification during the inference phase (where ground truth is unavailable), a complementary metric based on the model's predicted hand regions is necessary. Let $\mathcal{H}$ denote the set of pixels predicted by the model as belonging to the hand class. We define the average predictive entropy of predicted hand pixels ($\bar{E}_{\mathcal{H}}$) as follows:

$$\bar{E}_{\mathcal{H}} = \frac{1}{|\mathcal{H}|}\sum_{p\in \mathcal{H}} E(p), \tag{5}$$

where $|\mathcal{H}|$ represents the total number of pixels predicted as hand, and $E(p)$ is the predictive entropy of pixel p. Note that the set of predicted hand pixels ($\mathcal{H}$) differs from the set of ground truth hand pixels (H).

To leverage $\bar{E}_{\mathcal{H}}$ for uncertainty analysis, we establish a threshold using the validation dataset. Following the approach introduced in [44], we employ the Inter-Quartile Range (IQR) proximity rule for outlier detection to determine a reliable threshold (τ) as follows:

$$\tau = Q_3(\bar{E}_{\mathcal{H}}) + 1.5IQR(\bar{E}_{\mathcal{H}}), \tag{6}$$

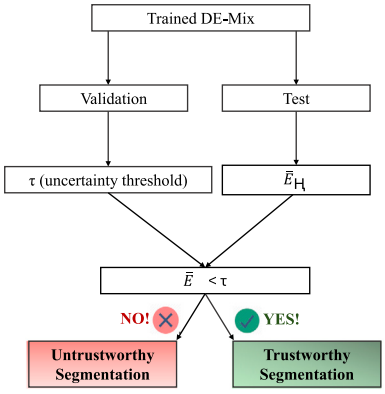


Fig. 8. Uncertainty analysis pipeline.

where $Q_3(\bar{E}\mathcal{H})$ denotes the third quartile, and $IQR(\bar{E}\mathcal{H})$ is the inter-quartile range (the difference between the third and first quartiles) computed over the validation dataset. During the test phase, predictions with average predictive entropy of predicted hand pixels ($\bar{E}_{\mathcal{H}}$) exceeding this threshold (τ) are flagged as untrustworthy, while those below it are considered trustworthy.

This pipeline thus serves as an OOD detection mechanism, allowing us to filter out predictions with high uncertainty. Ideally, this approach identifies most OOD scenarios as untrustworthy, thereby avoiding unreliable hand segmentation results. Fig. 8 illustrates the described uncertainty analysis pipeline.

4. Results and discussions

In this section, we evaluate the performance of the proposed deep ensemble models on human hand segmentation using metrics such as mIoU for segmentation accuracy and average predictive entropy for uncertainty quantification. The evaluations were performed under both ID and OOD conditions. We also present qualitative examples of some

Table 4
Prediction accuracy and predictive uncertainty of the DE-Mix model trained on each dataset using our test images with simple and cluttered backgrounds. Results are presented as mean and standard deviation (inside the parentheses) computed over five runs over 20 samples for each condition.

Training dataset	Test image condition	mIoU mean (SD)	$\bar{E}$ mean (SD)	$\bar{E}_h$ mean (SD)
EgoHands	One operator (O1) in simple background	0.382 (0.016)	0.114 (0.006)	0.195 (0.014)
	One operator (O1) in cluttered background	0.361 (0.021)	0.163 (0.010)	0.223 (0.019)
Ego2Hands	One operator (O1) in simple background	0.377 (0.020)	0.082 (0.003)	0.212 (0.016)
	One operator (O1) in cluttered background	0.318 (0.028)	0.140 (0.007)	0.275 (0.023)
HADR	One operator (O1) in simple background	0.419 (0.021)	0.095 (0.005)	0.229 (0.017)
	One operator (O1) in cluttered background	0.401 (0.030)	0.148 (0.009)	0.261 (0.024)
HAGS	One operator (O1) in simple background	0.496 (0.018)	0.158 (0.007)	0.186 (0.015)
	One operator (O1) in cluttered background	0.479 (0.022)	0.195 (0.011)	0.283 (0.025)

ID and OOD examples to discuss the segmentation performances in more detail.

4.1. Training and validation

For training, we selected annotated images from each dataset represented in Table 2. The data was split in a 9:1 ratio for training and validation respectively. It is important to note that the images in all four training datasets are in RGB format, except for HADR, which includes Depth information. For consistency, we utilized only the RGB format of the HADR dataset. All images were resized and reshaped to ensure compatibility with the generated test dataset.

The deep ensemble architecture (DE-Mix) was trained using the training data. Validation data was used to monitor the mIoU during training. We evaluated the performance of ensemble models with $K = 2, 4, 6, 8, 10$, and observed that model achieved an mIoU exceeding 0.80 on the validation set for $K \geq 4$ across all different training datasets. Based on these results, we selected $K = 4$ for all subsequent evaluations to balance performance and computational efficiency.

4.2. Quantitative results

We evaluated the trained models based on our test data, on egocentric images for models trained on EgoHands and Ego2Hands and on side-view images for models trained on HAGS and HADR-trained models. It is worth mentioning that all results presented are averaged over five runs for each of the 20 samples per condition, and the corresponding standard deviations are reported as well.

4.2.1. Background effect

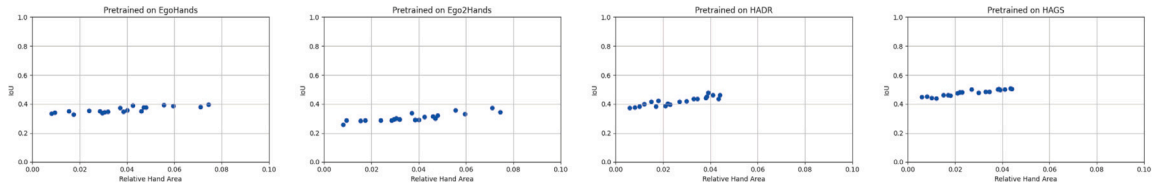
To consider the effect of background in images we compare the segmentation results of pre-trained models on our test images when one operator (O1) is interacting with robot in a simple background (without any other objects) and when the background is cluttered with a lot of industrial objects. The results regarding segmentation accuracy (mIoU) and entropy of predictions of entire image ($\bar{E}$) and hands specifically ($\bar{E}_h$) are listed in Table 4. Before presenting the results, it is important to clarify the performance metric ranges used in this study to facilitate interpretation. Since the models evaluated in this study are trained on datasets different from our custom-designed dataset, achieving significant mIoU values is not expected. Such performance drops have been reported in previous cross-dataset evaluations in the literature. For example, Sharma et al. [12], who introduced the HAGS dataset, reported that a UNet model trained on HADR achieved an mIoU ranging approximately from 0.30 to 0.45 when evaluated on HAGS under different conditions. Similarly, Lin et al. [14], who presented the Ego2Hands dataset, reported cross-dataset mIoUs around 0.26 to 0.33 between EgoHands and Ego2Hands, whereas training and testing within the same dataset resulted in mIoUs above 0.82. Regarding predictive entropy, its values range from 0 (complete certainty) to 1 (maximum uncertainty). Thus, lower entropy values indicate that the model predictions are more confident, whereas higher values reflect greater uncertainty in predictions.

As shown in Table 4, the segmentation accuracy (mIoU) for models trained on EgoHands and Ego2Hands is significantly lower compared to models trained on HADR and HAGS, regardless of background conditions. This discrepancy can be attributed to the non-industrial nature of the EgoHands and Ego2Hands datasets. The presence of industrial robots and workplace elements in the test images introduces challenges for these models, as their training data lacks such scenarios, reducing their generalizability to segment hands in industrial settings. In contrast, the model trained on HAGS outperforms the model trained on HADR, which may be due to the synthetic nature of the HADR dataset. Models trained on synthetic data often struggle with domain adaptation when applied to real-world scenarios. Additionally, the HADR dataset primarily contains images captured from a top-down view, whereas HAGS includes both top-down and side-view images, making it more similar to our test dataset and therefore better suited for segmentation in these conditions. Overall, segmentation accuracy decreased for all models when tested in cluttered and messy backgrounds compared to simple and minimalist backgrounds. This drop is more pronounced for models trained on egocentric datasets (EgoHands and Ego2Hands), likely due to the absence of industrial tools and environments in the backgrounds of their training data. Interestingly, model trained on HAGS was able to generalize well in the cluttered background while HAGS dataset does not cover a variety of industrial backgrounds. In general, the standard deviation of mIoU is relatively higher in cluttered backgrounds than in simple backgrounds for all trained models, indicating greater variability in model accuracy across different samples within the same condition when the scene is complex.

Examining the average predictive entropy of the entire image ($\bar{E}$) reveals that simpler backgrounds consistently result in lower entropy compared to cluttered backgrounds across all pre-trained models. This indicates that models are more confident in their predictions when the background is less complex while when the background is messy it increases the uncertainty of the model so that it struggles to classify each pixel to hand or background. Focusing on the pixels corresponding to hand regions, the average predictive entropy for hands ($\bar{E}_h$) is substantially higher than the overall entropy ($\bar{E}$), suggesting that models are inherently less confident when predicting hand pixels. Additionally, the standard deviation of $\bar{E}_h$ is higher than that of $\bar{E}$, meaning that the uncertainty associated with hand pixels varies more across samples than the overall uncertainty across all pixels within the same condition. This uncertainty becomes even more pronounced in cluttered backgrounds, likely because hands are often in close proximity to other objects, making it more challenging for models to confidently classify those pixels as hands.

4.2.2. Hand instance size effect

As discussed previously, the hand instance sizes within the training datasets exhibit different distributions (see Fig. 2). To examine how these size differences might affect segmentation accuracy, we investigated the relationship between hand size and the IoU score for the condition of one operator (O1) interacting in a cluttered background. For this purpose, we analyzed 20 samples from the test set in this condition, covering a relative hand area range of approximately 0.008



(a) EgoHands (b) Ego2Hands (c) HADR (d) HAGS

Fig. 9. IoU across different hand sizes for one operator (O1) in a cluttered background, for models trained on different datasets.

Table 5
Computed uncertainty thresholds (τ) for each model trained on different datasets.

Training dataset	Uncertainty threshold (τ)
EgoHands	0.221
Ego2Hands	0.234
HADR	0.279
HAGS	0.292

(smallest instance) to 0.075 (largest instance) for the egocentric view, and 0.008 to 0.044 for the side view, as hands appear smaller from the side viewpoint (Fig. 6). Fig. 9 shows the IoU results for these samples across models trained on each dataset. As observed from Fig. 9, all models, irrespective of the training dataset, tend to perform slightly better on larger hand instances, though this relationship is not strongly pronounced. Specifically, for models trained on EgoHands and Ego2Hands (Figs. 9(a) and 9(b)), despite differences in their original dataset distributions (Figs. 2(a) and 2(b)), larger hand instances achieved slightly higher IoU values—approximately 0.05 and 0.09 greater for the largest instances compared to the smallest, respectively. Similarly, models trained on HADR and HAGS datasets showed an IoU improvement of approximately 0.09 and 0.06, respectively, for larger hand instances. These observations suggest that while dataset distributions differed, all models exhibited similar, moderate sensitivity to hand instance size, with somewhat better segmentation performance observed for larger hand instances. Notably, a similar trend was reported previously in the HADR dataset study [15].

In conclusion, even though the distributions of hand sizes varied across the training datasets, the trained models did not exhibit significant bias towards specific hand sizes. This indicates that the models have a reasonable level of robustness to variations in hand instance sizes within the range examined.

4.2.3. ID and OOD data

As previously mentioned, we conducted uncertainty analysis to distinguish OOD data from ID data based on the threshold (τ) defined in Section 3.5. Table 5 reports the specific thresholds computed for each model trained on EgoHands, Ego2Hands, HADR, and HAGS datasets. Following the pipeline illustrated in Fig. 8, test samples whose average predictive entropy of predicted hand pixels ($\bar{E}_{\mathcal{H}}$) exceeds these thresholds are flagged as untrustworthy predictions (potentially OOD), enabling more reliable and interpretable segmentation outcomes.

The segmentation performance (mIoU), predictive entropy, and percentage of detected as untrustworthy for the deep ensemble model (DE-Mix) trained on four training datasets are listed in Table 6 for both in-distribution (ID) and out-of-distribution (OOD) data. Notably, all test conditions involve messy and cluttered backgrounds with industrial tools, ensuring realistic and challenging evaluation scenarios. To provide a structured analysis, the discussion is divided into two parts: segmentation accuracy (mIoU) and predictive entropy.

Accuracy of segmentation (mIoU): As can be seen from Table 6, the trained model has better segmentation accuracy (mIoU) for the conditions it encountered during training (ID data) compared to the scenarios it faced without prior exposure (OOD data) across all four training datasets. Aside from presenting their mean and standard deviation in Table 6, we further analyzed the IoU of each condition (see the appendix) ensuring their distributions are meaningfully different. The model trained on EgoHands when segment hands of one or two human operators in the scene (O1 and O2), maintains its accuracy for two operators since this condition was present during training. In contrast, the model trained on Ego2Hands experiences a significant accuracy drop in the same scenario since it has not faced two humans in its training (OOD data). In other OOD scenarios (e.g., two operators with and without gloves), models trained on these egocentric datasets perform poorly, with mIoU values below 0.2. Particularly for rare gestures (RG), these models fail to segment even a small portion of the operators' hands (mIoU less than 0.1). Similarly, for motion-blurred images (MBN), representing aleatoric uncertainty, these models achieve an mIoU around 0.1, effectively failing to segment hands. On the other hand, models trained on HADR and HAGS datasets show significantly better segmentation accuracy in both ID and OOD data. This performance can be attributed to their industrial context, which closely matches the testing conditions in this study. The model trained on HAGS performs better than the one trained on HADR, likely because HAGS includes gloved hands (GH), contains both top-down and side-view perspectives and its data is realistic (not synthetic), which are similar to our test dataset. In O2 scenarios (two operators), both HAGS- and HADR-trained models experience an accuracy drop, as these conditions are OOD for their training sets. However, the HAGS-trained model achieves better mIoU (0.35) compared to the HADR-trained model. In rare gesture (RG) and motion-blurred (MBN) scenarios, both models struggle, but the HAGS-trained model performs slightly better (mIoU 0.23 for RG, 0.28 for MBN) compared to HADR (mIoU 0.12 for RG, 0.24 for MBN).

Predictive Entropy and uncertainty analysis ($\bar{E}$, $\bar{E}_h$, $\bar{E}_{\mathcal{H}}$, and untrustworthy percentage): From Table 6, it can be observed that the average predictive entropy for the entire image ($\bar{E}$) remains relatively stable between ID and OOD data with a low standard deviation over different runs across samples in each condition. This stability contrasts with the significant change observed in Table 4 when comparing simple and cluttered backgrounds. This is likely because hand regions constitute a small proportion of the entire image, so changes in hand conditions do not significantly affect the overall uncertainty of the image. However, when examining the predictive entropy for hand pixels ($\bar{E}_h$), a good model is expected to exhibit higher uncertainty for OOD data compared to ID data. Models trained on EgoHands and Ego2Hands do not consistently follow this pattern. In these models, the predictive entropy of predicted hand pixels ($\bar{E}_{\mathcal{H}}$) of OOD data is also closer to ID data, and the percentage of OOD data flagged as untrustworthy is relatively low in some conditions. Despite this, the uncertainty-based filtering is still useful since approximately none of ID data was flagged incorrectly as untrustworthy (since ($\bar{E}_{\mathcal{H}}$) of ID data was lower than thresholds obtained from Table 5). Also, in some cases (such as two operators with gloves), a considerable portion of the OOD data was detected as untrustworthy. This analysis is helpful for reliable segmentation, such that the OOD and ID can be better separated in the inference phase and potentially discarding results likely to be erroneous.

However, models trained on HAGS and HADR report higher both ($\bar{E}_h$) and $\bar{E}_{\mathcal{H}}$ in OOD conditions, indicating appropriate recognition of unfamiliar scenarios as uncertain.

Table 6
Segmentation performance, predictive uncertainty, and uncertainty analysis results of models trained on each dataset using our test data in ID and OOD conditions. Results are presented as mean and standard deviation (inside the parentheses) computed over five runs over 20 samples for each condition.

Training dataset	ID/OOD	Image condition in test phase	mIoU mean (SD)	$\bar{E}$ mean (SD)	$\bar{E}_h$ mean (SD)	$\bar{E}_{\mathcal{H}}$ mean (SD)	Untrustworthy mean (SD)
EgoHands	ID	One operator (O1)	0.361 (0.021)	0.163 (0.010)	0.223 (0.019)	0.177 (0.019)	0% (0%)
		Two operators (O2)	0.306 (0.023)	0.172 (0.012)	0.278 (0.022)	0.192 (0.018)	8% (7%)
	OOD	One operator (O1) with gloves (GH)	0.184 (0.026)	0.157 (0.013)	0.211 (0.021)	0.204 (0.023)	20% (8%)
		Two operators (O2) with gloves (GH)	0.115 (0.030)	0.174 (0.015)	0.215 (0.024)	0.212 (0.022)	41% (8%)
		Two operators (O2) having rare gestures (RG)	0.052 (0.029)	0.168 (0.016)	0.231 (0.019)	0.225 (0.021)	51% (8%)
		Two operators (O2) with gloves (GH) having rare gestures (RG)	0.058 (0.031)	0.170 (0.015)	0.164 (0.020)	0.198 (0.025)	18% (17%)
		Images with motion blurred noise (MBN)	0.103 (0.032)	0.154 (0.011)	0.126 (0.015)	0.178 (0.023)	4% (2%)
Ego2Hands	ID	One operator (O1)	0.318 (0.028)	0.140 (0.007)	0.275 (0.023)	0.205 (0.019)	2% (4%)
	OOD	Two operators (O2)	0.202 (0.033)	0.153 (0.010)	0.247 (0.025)	0.223 (0.022)	27% (9%)
		One operator (O1) with gloves (GH)	0.171 (0.037)	0.134 (0.008)	0.198 (0.021)	0.209 (0.024)	11% (5%)
		Two operators (O2) with gloves (GH)	0.105 (0.034)	0.165 (0.013)	0.250 (0.024)	0.243 (0.027)	65% (10%)
		Two operators (O2) having rare gestures (RG)	0.069 (0.033)	0.146 (0.011)	0.212 (0.019)	0.225 (0.025)	37% (11%)
		Two operators (O2) with gloves (GH) having rare gestures (RG)	0.033 (0.020)	0.162 (0.016)	0.298 (0.028)	0.251 (0.029)	71% (4%)
		Images with motion blurred noise (MBN)	0.115 (0.034)	0.149 (0.013)	0.165 (0.019)	0.192 (0.023)	3% (2%)
HADR	ID	One operator (O1)	0.419 (0.021)	0.148 (0.009)	0.261 (0.024)	0.219 (0.025)	0% (0%)
	OOD	Two operators (O2)	0.349 (0.032)	0.155 (0.016)	0.312 (0.028)	0.288 (0.031)	61% (10%)
		One operator (O1) with gloves (GH)	0.279 (0.031)	0.131 (0.010)	0.329 (0.030)	0.273 (0.029)	45% (8%)
		Two operators (O2) with gloves (GH)	0.201 (0.036)	0.160 (0.017)	0.412 (0.033)	0.313 (0.030)	92% (7%)
		Two operators (O2) having rare gestures (RG)	0.148 (0.031)	0.152 (0.016)	0.403 (0.035)	0.361 (0.033)	99% (2%)
		Two operators (O2) with gloves (GH) having rare gestures (RG)	0.123 (0.028)	0.150 (0.014)	0.365 (0.026)	0.307 (0.027)	77% (10%)
		Images with motion blurred noise (MBN)	0.239 (0.029)	0.135 (0.011)	0.287 (0.023)	0.251 (0.025)	17% (8%)
HAGS	ID	One operator (O1)	0.479 (0.022)	0.195 (0.011)	0.283 (0.025)	0.226 (0.017)	0% (0%)
		One operator (O1) with gloves (GH)	0.465 (0.025)	0.182 (0.015)	0.272 (0.022)	0.212 (0.020)	0% (0%)
	OOD	Two operators (O2)	0.362 (0.029)	0.204 (0.019)	0.348 (0.029)	0.310 (0.031)	75% (7%)
		Two operators (O2) with gloves (GH)	0.346 (0.031)	0.199 (0.020)	0.302 (0.031)	0.308 (0.026)	65% (9%)
		Two operators (O2) having rare gestures (RG)	0.238 (0.032)	0.196 (0.018)	0.417 (0.037)	0.391 (0.038)	98% (3%)
		Two operators (O2) with gloves (GH) having rare gestures (RG)	0.229 (0.033)	0.201 (0.017)	0.404 (0.034)	0.356 (0.033)	97% (2%)
		Images with motion blurred noise (MBN)	0.278 (0.030)	0.193 (0.015)	0.292 (0.024)	0.288 (0.029)	50% (7%)

Accordingly, the untrustworthy percentage is high for OOD data (with low standard deviations across different runs) while remaining negligible for ID data. This pattern results in a more reliable segmentation pipeline where OOD and ID data can be separated at inference time, and segmentation results can be trusted with greater confidence. This is especially pronounced for the model trained on HAGS, which flags between 50% and 98% of OOD data as untrustworthy, thus achieving highly reliable performance under data distributional shift.

The superiority of models trained on HAGS and HADR is clearly evident compared to those trained on EgoHands and Ego2Hands in terms of segmentation accuracy and uncertainty analysis. This can be attributed to the fact that the images in HAGS and HADR datasets are more aligned with our testing conditions, as both datasets are specifically designed for industrial human–robot interaction scenarios. In contrast, the images in EgoHands and Ego2Hands primarily depict hands in general, non-industrial activities, making them less relevant for the industrial context of our evaluation.

It is important to note that the results reported above are based on models evaluated directly without fine-tuning on the test dataset. This deliberate choice ensures that the evaluation highlights the models' generalization capabilities in novel, realistic industrial conditions. However, this approach inherently results in lower segmentation accuracy (mIoU) compared to models that undergo fine-tuning on similar test data.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU, an AMD Ryzen 9 processor, and 64 GB of memory. This setup provided the computational power required for training and evaluating the deep ensemble model (DE-Mix). With this configuration, the DE-Mix model (with $K = 4$ base learners) achieved a segmentation speed of 38 frames per second, demonstrating its capability for real-time performance in practical applications.

4.3. Qualitative results

The segmentation results for both ID and OOD images captured from egocentric and side cameras are presented in this section.

Fig. 10 shows an egocentric image of ID data, depicting one operator (O1) interacting with the environment. The ground truth and predicted segmentation masks from models trained on EgoHands and Ego2Hands are displayed. The model trained on EgoHands is able to segment some pixels of both hands of the operator but mistakenly identifies part of the wooden handle of the hammer as hand pixels. In contrast, the model trained on Ego2Hands segments only a small portion of one hand and mistakenly assigns hand labels to the wooden handle of the hammer and another nearby tool. These results indicate the limited accuracy of models trained on these datasets even dealing with ID data. In both cases, the models did not flag these segmentations as untrustworthy, since the predictive entropy of the predicted hand pixels ($\bar{E}_{\mathcal{H}}$) remained below the corresponding threshold (τ). This suggests that the uncertainty-based filter does not mistakenly reject segmentation on ID data, demonstrating its reliability in retaining trustworthy ID predictions.

Fig. 11 depicts a side-view image of the same condition shown in Fig. 10, with segmentation results from models trained on HADR and HAGS. The model trained on HADR segments some pixels of one hand while misclassifying parts of the tool carried by the operator as hand pixels. Conversely, the model trained on HAGS demonstrates better performance by detecting portions of both hands, even though the operator's left hand is occluded by the industrial robot. This highlights the higher accuracy of the HAGS-trained model compared to the HADR-trained model in ID scenarios. Consistent with the previous example, the predicted segmentations in this side-view ID scenario were correctly not flagged as untrustworthy.

Fig. 12 illustrates an OOD image featuring two operators (O2) wearing gloves (GH) and performing rare gestures (RG). The segmentation masks from the model trained on EgoHands reveal segmenting only a few pixels of the operators' forearms, which are not considered part of the hand since training and testing data define hands as the region from the fingertips to the wrist. Additionally, the model incorrectly identifies parts of the robotic arm and the wooden hammer handle as hand pixels. The model trained on Ego2Hands, however, fails to segment any human

Fig. 10. (a) an ID data in the egocentric view, (b) the segmentation ground truth mask, (c) segmented by DE-Mix trained on EgoHands, (d) segmented by DE-Mix trained on Ego2Hands.

Fig. 11. (a) an ID data in the side view, (b) the segmentation ground truth mask, (c) segmented by DE-Mix trained on HADR, (d) segmented by DE-Mix trained on HAGS.

Fig. 12. (a) an OOD data of two human operators O2 with gloves (GH) having rare gestures(RG) in the he egocentric view, (b) the segmentation ground truth mask, (c) segmented by DE-Mix trained on EgoHands, (d) segmented by DE-Mix trained on Ego2Hands.

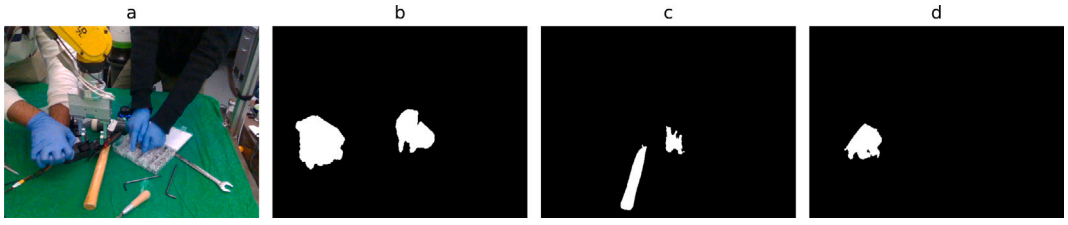


Fig. 13. (a) an OOD data of two human operators O2 with gloves (GH) having rare gestures(RG) in the side view, (b) the segmentation ground truth mask, (c) segmented by DE-Mix trained on HADR, (d) segmented by DE-Mix trained on HAGS.

hand pixels while erroneously assigning hand labels to parts of the robotic arm and tools. These observations explain the near-zero mIoU values reported for these models when dealing with these kinds of OOD data, as shown in Table 6. Interestingly, for this challenging OOD case, the model trained on Ego2Hands correctly flagged the segmentation as untrustworthy. Conversely, the model trained on EgoHands failed to flag the segmentation as untrustworthy.

Fig. 13 displays the side-view perspective of the same OOD scenario shown in Fig. 12. The model trained on HADR correctly segments a small portion of one operator's hand but misclassifies the wooden hammer handle as hand pixels. The model trained on HAGS, however, accurately segments portions of one operator's hand without incorrectly assigning other pixels as hands, although it fails to detect the other operator's hands. These results align with the lower mIoU values observed for OOD data compared to ID data, as reported in Table 6. For both models in this OOD scenario, the resulting segmentations were correctly identified as untrustworthy during inference, as the predictive entropy for the predicted hand regions exceeded the threshold.

5. Conclusions and future works

This study investigated the performance of a deep ensemble model, DE-Mix, composed of UNet and RefineNet, for human hand segmentation in industrial human–robot interaction scenarios. Using four datasets EgoHands, Ego2Hands, HADR, and HAGS we evaluated the generalization capabilities of models trained on these datasets under diverse and challenging conditions, including both in-distribution (ID) and out-of-distribution (OOD) scenarios. The results underscored the importance of domain-specific datasets, with models trained on HAGS and HADR consistently outperforming those trained on EgoHands and Ego2Hands in segmentation accuracy and predictive uncertainty metrics. In addition, we introduced an uncertainty analysis pipeline to flag OOD data as untrustworthy during test. This provides a mechanism for identifying unreliable segmentations in practical deployment.

Although OOD conditions posed significant challenges for all models in achieving accurate hand segmentation, those trained on industrial datasets demonstrated greater robustness in handling scenarios such as gloved hands, rare gestures, and motion blur. In contrast, models trained on non-industrial datasets struggled to generalize, often failing to accurately segment hands or misclassifying non-hand regions. Importantly, the proposed uncertainty analysis pipeline was effective in filtering out unreliable segmentations. A substantial proportion of OOD data was correctly flagged as untrustworthy, particularly for models trained on HADR and HAGS. At the same time, the pipeline maintained high reliability for ID data, as almost no ID samples were incorrectly rejected across all models, and even models trained on EgoHands and Ego2Hands benefited by correctly flagging certain OOD cases. This

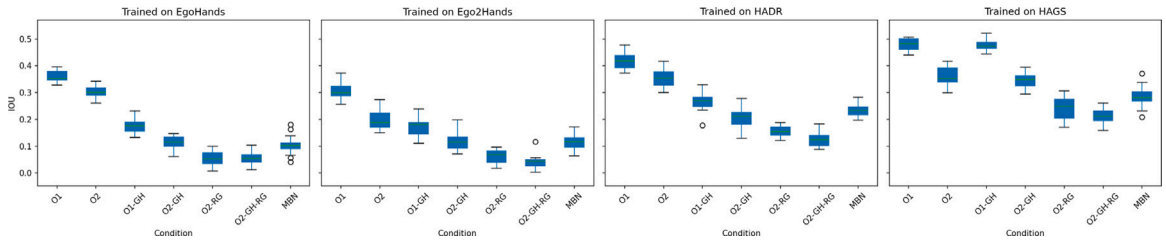


Fig. A.1. Distribution of IoU values across 20 samples for each test condition with respect to Table 6.

indicates that the entropy-based filtering enhances the trustworthiness of segmentation outputs under challenging real-world conditions.

The complex and realistic nature of the test dataset characterized by cluttered backgrounds, industrial tools, and multiple human hands highlighted the inherent challenges of deploying segmentation models in dynamic, real-world environments. While segmentation accuracy was relatively low due to the absence of fine-tuning on the test data, this limitation emphasized the challenges of generalizing to novel industrial conditions, reinforcing the need for more diverse datasets and adaptive techniques.

Future research could focus on fine-tuning pre-trained models with industrial datasets to enhance their applicability to specific scenarios. Expanding the diversity of training datasets by data augmentation and also incorporating rare gestures, motion blur, and occluded hands could be another option. Although constructing such comprehensive datasets is a resource-intensive and time-consuming and inclusion of certain conditions (particularly rare gestures) comprehensively into datasets remains challenging due to their inherent variability and unpredictability. We also restrict our experience to two human operators and also two cameras, i.e. egocentric and side camera, extending future work to include more complex scenarios, such as multiple operators in the scene of interaction and different camera placements by considering different angles and distances from participants, would provide further insights. Additionally, integrating multimodal data, such as depth, may further enhance segmentation accuracy under challenging conditions.

Exploration of advanced model architectures, such as Transformers, offers promising potential to improve spatial understanding and generalization to OOD scenarios. Furthermore, optimizing lightweight models for real-time inference would facilitate deployment in industrial environments, where speed and efficiency are critical.

CRediT authorship contribution statement

Reza Jalayer: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yuxin Chen:** Writing – review & editing, Visualization, Investigation, Data curation. **Masoud Jalayer:** Writing – review & editing, Visualization, Validation, Software, Methodology, Conceptualization. **Carlotta Orsenigo:** Writing – review & editing, Supervision, Resources, Project administration. **Masayoshi Tomizuka:** Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Fig. A.1 shows the distribution of IoU values across 20 samples for each test condition (both ID and OOD conditions listed in Table 6), evaluated using the four pre-trained models. As evident, the distributions vary across conditions for each model. To statistically validate this observation, we applied the Kruskal–Wallis test separately for each pre-trained model to assess whether segmentation performance differed significantly across conditions. The test results were statistically significant in all cases (p-value $< 1.3\times10^{-23}$), indicating substantial variation in IoU distributions. Therefore, the null hypothesis, stating that there is no significant difference in model performance across test conditions, was rejected in each case.

Data availability

Data will be made available on request.

References

[1] Grau A, Indri M, Bello LL, Sauter T. Robots in industry: The past, present, and future of a growing collaboration with humans. IEEE Ind Electron Mag 2020;15(1):50–61.
[2] Tantawi KH, Sokolov A, Tantawi O. Advances in industrial robotics: From industry 3.0 automation to industry 4.0 collaboration. In: 2019 4th technology innovation management and engineering science international conference (TIMES-ICON). IEEE; 2019, p. 1–4.
[3] Kim J. Collision detection and reaction for a collaborative robot with sensorless admittance control. Mechatronics 2022;84:102811.
[4] Benmessabih T, Slama R, Havard V, Baudry D. Online human motion analysis in industrial context: A review. Eng Appl Artif Intell 2024;131:107850.
[5] Alves J, Lima TM, Gaspar PD. Is industry 5.0 a human-centred approach? a systematic review. Processes 2023;11(1):193.
[6] Qi J, Ma L, Cui Z, Yu Y. Computer vision-based hand gesture recognition for human-robot interaction: a review. Complex Intell Syst 2024;10(1):1581–606.
[7] Gao Q, Chen Y, Ju Z, Liang Y. Dynamic hand gesture recognition based on 3D hand pose estimation for human–robot interaction. IEEE Sensors J 2021;22(18):17421–30.
[8] Huang H, Liang Z, Sun F, Dong M, et al. Virtual interaction and manipulation control of a hexacopter through hand gesture recognition from a data glove. Robotica 2022;40(12):4375–87.
[9] Tamantini C, Cordella F, Lauretti C, Zollo L. The WGD—A dataset of assembly line working gestures for ergonomic analysis and work-related injuries prevention. Sensors 2021;21(22):7600.
[10] Yun L, Lifeng Z, Shujun Z. A hand gesture recognition method based on multi-feature fusion and template matching. Procedia Eng 2012;29:1678–84.
[11] Taran O, Manzone DM, Zariffa J. Benchmarking 2D egocentric hand pose datasets. 2024, arXiv preprint arXiv:2409.07337.
[12] Sharma S, Huang M, Nair S, Wen A, Petlowany C, Moore J, Wanna S, Pryor M. The collection of a human robot collaboration dataset for cooperative assembly in glovebox environments. 2024, arXiv preprint arXiv:2407.14649.
[13] Bambach S, Lee S, Crandall DJ, Yu C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 1949–57.
[14] Lin F, Price B, Martinez T. Ego2hands: A dataset for egocentric two-hand segmentation and detection. 2020, arXiv preprint arXiv:2011.07252.
[15] Grushko S, Vysockỳ A, Chlebek J, Prokop P. HaDR: Applying domain randomization for generating synthetic multimodal dataset for hand instance segmentation in cluttered industrial environments. 2023, arXiv preprint arXiv:2304.05826.
[16] Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, Fieguth P, Cao X, Khosravi A, Acharya UR, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion 2021;76:243–97.
[17] Redmon J. You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[18] Chen Y, Wei W, Xiao W. Human-computer interaction control of snake-like robot based on gesture recognition. In: Proceedings of the 2019 4th international conference on automation, control and robotics engineering. 2019, p. 1–6.

[19] Panteleris P, Oikonomidis I, Argyros A. Using a single rgb frame for real time 3d hand pose estimation in the wild. In: 2018 IEEE winter conference on applications of computer vision. WACV, IEEE; 2018, p. 436–45.
[20] Cao Z, Simon T, Wei S-E, Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 7291–9.
[21] Zhang F, Bazarevsky V, Vakunov A, Tkachenka A, Sung G, Chang C-L, Grundmann M. Mediapipe hands: On-device real-time hand tracking. 2020, arXiv preprint arXiv:2006.10214.
[22] Mazhar O, Navarro B, Ramdani S, Passama R, Cherubini A. A real-time human-robot interaction framework with robust background invariant hand gesture detection. Robot Comput-Integr Manuf 2019;60:34–48.
[23] Docekal J, Rozlivek J, Matas J, Hoffmann M. Human keypoint detection for close proximity human-robot interaction. In: 2022 IEEE-RAS 21st international conference on humanoid robots (humanoids). IEEE; 2022, p. 450–7.
[24] Dutta HPJ, Bhuyan MK, Neog DR, MacDorman KF, Laskar RH. Efficient hand segmentation for rehabilitation tasks using a convolution neural network with attention. Expert Syst Appl 2023;234:121046.
[25] Bandini A, Zariffa J. Analysis of the hands in egocentric vision: A survey. IEEE Trans Pattern Anal Mach Intell 2020;45(6):6846–66.
[26] Rastgoo R, Kiani K, Escalera S. Hand sign language recognition using multi-view hand skeleton. Expert Syst Appl 2020;150:113336.
[27] Wu Y, Liu Y, Wang J. Real-time hand-object occlusion for augmented reality using hand segmentation and depth correction. In: 2023 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops. VRW, IEEE; 2023, p. 631–2.
[28] Hassan EK, Jamila Harbi S. 3D hand pose and shape estimation from single RGB image for augmented reality. J Intell Syst Internet Things 2023;10(2):90–101.
[29] Sahoo JP, Sahoo SP, Ari S, Patra SK. Hand gesture recognition using densely connected deep residual network and channel attention module for mobile robot control. IEEE Trans Instrum Meas 2023;72:1–11.
[30] Xu J, Li J, Zhang S, Xie C, Dong J. Skeleton guided conflict-free hand gesture recognition for robot control. In: 2020 11th international conference on awareness science and technology. ICAST, IEEE; 2020, p. 1–6.
[31] Gao Q, Liu J, Ju Z, Zhang X. Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation. IEEE Trans Ind Electron 2019;66(12):9663–72.
[32] van Amsterdam B, Clarkson MJ, Stoyanov D. Gesture recognition in robotic surgery: a review. IEEE Trans Biomed Eng 2021;68(6).
[33] Sajedi S, Liu W, Eltouny K, Behdad S, Zheng M, Liang X. Uncertainty-assisted image-processing for human-robot close collaboration. IEEE Robot Autom Lett 2022;7(2):4236–43.
[34] Vysocky A, Grushko S, Spurny T, Pastor R, Kot T. Generating synthetic depth image dataset for industrial applications of hand localization. IEEE Access 2022;10:99734–44.
[35] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B, Snoek J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. Adv Neural Inf Process Syst 2019;32.
[36] Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label studio: Data labeling software. 2022, 2020, Open source software available from https://github.com/heartexlabs/label-studio.
[37] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–mICCAI 2015: 18th international conference, munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer; 2015, p. 234–41.
[38] Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1925–34.
[39] Dutta HPJ, Sarma D, Bhuyan MK, Laskar RH. Semantic segmentation based hand gesture recognition using deep neural networks. In: 2020 national conference on communications. NCC, IEEE; 2020, p. 1–6.
[40] Urooj A, Borji A. Analysis of hand segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 4710–9.
[41] Zhang C, Han D, Qiao Y, Kim JU, Bae S-H, Lee S, Hong CS. Faster segment anything: Towards lightweight sam for mobile applications. 2023, arXiv preprint arXiv:2306.14289.
[42] Yu C, Gao C, Wang J, Yu G, Shen C, Sang N. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis 2021;129:3051–68.
[43] Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: A review. Eng Appl Artif Intell 2022;115:105151.
[44] Han T, Li Y-F. Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles. Reliab Eng Syst Saf 2022;226:108648.