# Solid and Effective Upper Limb Segmentation in Egocentric Vision

Monica Gruosso
University of Basilicata
Potenza, Italy
monica.gruosso@unibas.it

Nicola Capece
University of Basilicata
Potenza, Italy
nicola.capece@unibas.it

Ugo Erra
University of Basilicata
Potenza, Italy
ugo.erra@unibas.it

**Figure 1: Real-life photos captured in egocentric vision (top row) and segmentation overlay images (bottom row) obtained by overlapping the input and prediction of our best model based on DeepLabv3+ with Xception-65 network backbone. Accurate segmentation was achieved for both naked and clothed upper limbs in different lighting conditions, skin tone, occlusions, hand poses, user/camera movements, indoor and outdoor scenarios.**

## ABSTRACT

Upper limb segmentation in egocentric vision is a challenging and nearly unexplored task that extends the well-known hand localization problem and can be crucial for a realistic representation of users' limbs in immersive and interactive environments, such as VR/MR applications designed for web browsers that are a general-purpose solution suitable for any device. Existing hand and arm segmentation approaches require a large amount of well-annotated data. Then different annotation techniques were designed, and several datasets were created. Such datasets are often limited to synthetic and semi-synthetic data that do not include the whole limb and differ significantly from real data, leading to poor performance in many realistic cases. To overcome the limitations of previous methods and the challenges inherent in both egocentric vision and segmentation, we trained several segmentation networks based on the state-of-the-art DeepLabv3+ model, collecting a large-scale comprehensive dataset. It consists of 46 thousand real-life and well-labeled RGB images with a great variety of skin colors, clothes, occlusions, and lighting conditions. In particular, we carefully selected the best data from existing datasets and added our EgoCam dataset, which includes new images with accurate labels. Finally, we extensively evaluated the trained networks in unconstrained real-world environments to find the best model configuration for this task, achieving promising and remarkable results in diverse scenarios. The code, the collected egocentric upper limb segmentation dataset, and a video demo of our work will be available on the project page[1].

## CCS CONCEPTS

• **Computing methodologies** → **Mixed / augmented reality**; **Virtual reality**; **Image segmentation**; **Neural networks**; Perception; *Image processing*.

## KEYWORDS

Semantic Segmentation, Neural Networks, Computer Vision, Virtual Reality, Augmented Reality, Web MR/VR, Dataset, Egocentric Vision

**ACM Reference Format:**
Monica Gruosso, Nicola Capece, and Ugo Erra. 2021. Solid and Effective Upper Limb Segmentation in Egocentric Vision. In *The 26th International Conference on 3D Web Technology (Web3D '21), November 8–12, 2021, Pisa, Italy.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485444.3495179

## 1 INTRODUCTION

With the widespread availability of wearable RGB cameras and head-mounted displays, systems for analyzing and detecting hands in a first-person perspective, called egocentric or first-person vision

[1]Data and code available for research purposes and the video demo of our work can be found at the following link: http://graphics.unibas.it/www/EgoUpperLimbSeg/index.md.html


*Web3D '21, November 8–12, 2021, Pisa, Italy*

ACM ISBN 978-1-4503-9095-8/21/11…$15.00
https://doi.org/10.1145/3485444.3495179

(FPV), have increasingly developed [Betancourt et al. 2015]. It has led to the growing development of hand-based approaches in FPV in several areas, such as healthcare monitoring [Likitlersuang et al. 2019], gesture and sign language recognition [Zamora-Mora and Chacón-Rivas 2019], human-computer and human-robot interaction [Caggianese et al. 2016; Pathi et al. 2019], virtual, augmented, and mixed reality [Caggianese et al. 2015; Gonzalez-Sosa et al. 2020]. Hand segmentation is the most demanding hand localization task and is used as a pre-processing step in many applications since it allows identifying hand regions with high accuracy and distinguishing the hands from the background and objects [Bandini and Zariffa 2020]. In this study, we extended the hand segmentation task focusing on the upper limb in egocentric vision in unconstrained real-life environments. Understanding where the whole upper limb is, with a certain precision at pixel-level, can be useful especially in real-world scenarios where not only the hand but the rest of the upper limb is framed, for example, using cameras with a wide field of view. Furthermore, upper limb segmentation can be beneficial for constructing a realistic solution to represent users in virtual environments (VEs) visualized through head-mounted displays (HMDs). Although VR allows users to be involved in experiences that are difficult or impossible to achieve in the real world, many applications are not fully accessible to everyone and anywhere. Indeed, they are often coupled to specific platforms and require software and libraries installation. In recent years, Web3D-based solutions have been increasingly developed, providing an execution environment running directly in the browser and resulting in easy setup and high cross-platform portability [Chittaro and Serra 2004; de Paiva Guimarães et al. 2018]. Two interesting applications consist of VR chatrooms and Web3D virtual conferences, in which each user can interact with 3D virtual objects and other participants immersed in a Web3D virtual environment (Web3D VE) [Hok et al. 2020]. Such VR applications usually employ virtual avatars that are often limited to hand representation or look artificial [Gonzalez-Sosa et al. 2020]. Instead, the proposed approach can allow users to see their real whole limbs in the virtual scene and interact with virtual objects using their own hands, improving the user's sense of presence and embodiment [Kilteni et al. 2012].

Contrary to many state-of-the-art approaches, which segmented only the hand up to the wrist [Bambach et al. 2015; Urooj and Borji 2018] or bare arms [Li and Kitani 2013; Lin and Martinez 2020; Wang et al. 2019], we were hence interested in the whole upper limb, also taking into account the clothes. Moreover, we considered hand-to-hand and hand-to-object occlusions, treating the limb as foreground and the objects as part of the background. The main problem we faced was the lack of well-annotated realistic RGB images that included all the cases we wanted to analyze. Most existing datasets contain insufficient variety and a limited number of annotated images (e.g., a few hundred or thousand), making them unsuitable for training deep neural networks [Bandini and Zariffa 2020]. Larger datasets consist of easily labeled computer-generated data that differs from the realistic domain in terms of chromaticity, lighting conditions, shadows, and overall appearance. Such discrepancies may not allow the model to generalize on real-life cases or different datasets. Therefore, we collected about 46 thousand varied RGB images with accurate labels, which may enable a deep neural network to learn a wide range of realistic activities and achieve acceptable to remarkable results in unknown and unpredictable cases, without any fine-tuning or new training. In particular, the data includes *i)* the best images and labels from the well-known EDSH [Li and Kitani 2013] and TEgO [Lee and Kacorri 2019] datasets, *ii)* our manually labeled dataset, called EgoCam, obtained using two different cameras in egocentric vision.

To achieve our goal, we used the collected upper limb segmentation dataset to train and test several networks based on the DeepLabv3+ architecture [Chen et al. 2018], which is the state-of-the-art deep convolutional networks for semantic segmentation that obtained remarkable results on several benchmark datasets. Finally, we conducted quantitative and qualitative assessments to verify the effectiveness of our dataset and evaluate the best network configuration, achieving promising and notable results for both simple and complex/cluttered background scenes, different lighting conditions, hand poses, occlusions, and dynamic camera positions combined with the user's motion, without the need for domain adaptation. To the best of our knowledge, the proposed work is the first to evaluate the effectiveness of a deep learning (DL) model for upper limb segmentation in unconstrained realistic cases and collecting vast and varied real-life well-annotated images.

The remainder of this paper is structured as follows: Section 2 presents an overview of the related works; Section 3 analyzes the existing dataset, describes the requirements for selecting the best data and provides detailed information about our EgoCam dataset; Section 4 presents the considered networks, configuration parameters and training phase; Section 5 illustrates the obtained results and evaluations conducted; Section 6 presents our final considerations and future directions.

## 2 RELATED WORK

In this Section, we provide an overview of the hand/arm based approaches, discuss the strengths and weaknesses of synthetic and semi-synthetic datasets, and carefully analyze existing real datasets containing RGB images of hands and arms captured in egocentric vision. Finally, the use of deep learning techniques in Web3D-based applications was analyzed.

*Hand and Arm Based Methods.* Hands/arms based approaches can be divided into localization and interpretation methods [Bandini and Zariffa 2020]. The former include all studies interested in identifying the presence of hands and arms in images (detection) [Narasimhaswamy et al. 2019], classifying left/right hands (hand identification) [Betancourt et al. 2017], identifying the hand/arm pixel regions (segmentation) [Gonzalez-Sosa et al. 2020], and deriving the position of the hand joints (hand pose estimation or hand tracking) [Capece et al. 2020; Gruosso et al. 2020; Zimmermann and Brox 2017]. The latter collect all those methods that can deduce high-level information starting from those obtained by the localization methods, such as the identification of gestures, activities, and interactions [Cai et al. 2017; Nguyen et al. 2018].

One of the first hand segmentation approaches based on deep learning was proposed by A. Betancourt *et al.* [Betancourt et al. 2017], who extended traditional methods by introducing an intermediate hand-identification step to detect right and left hands using a Maxwell distribution of angle and position. In this way,

the authors demonstrated that hand segmentation in egocentric videos can benefit from left/right discrimination. Y. Li *et al.* [Li et al. 2019] introduced an unsupervised and semi-supervised hand segmentation method for egocentric images starting from a fully convolutional neural network (FCNN) pre-trained using pixel-level annotations dataset, and subsequently that FCNN was re-trained using unlabeled dataset with optimized mask annotations. Another approach for hand segmentation was designed by M. Cai *et al.* [Cai et al. 2020]. It consisted of a model adaptation framework based on Bayesian CNN to deal with the typical generalization problem that affects this type of task and the scarcity of large well-annotated datasets. Conversely, some recent methods collected a large amount of low-cost artificial data to train deep learning models, as proposed by Gonzalez *et al.* [Gonzalez-Sosa et al. 2020]. In detail, they trained a deep neural network using their own semi-synthetic egocentric arm segmentation dataset, which consisted of more than 10 thousand composited images. Although their model achieved interesting results, several false positive and segmentation errors in color similarities between background and foreground were found.

To the best of our knowledge, the proposed work is the first to design a deep learning approach for the segmentation of the whole upper limb in unconstrained environments and different lighting conditions, skin tone, clothes, and occlusions.

*Hand and Arm Datasets.* Since obtaining ground-truth segmentation masks is very labor-intensive, time-consuming, and sometimes impractical, synthetic and semi-synthetic data have been widely collected, allowing to gather many images and get labels quickly and with little effort [Gruosso et al. 2021; Lin and Martinez 2020; Mueller et al. 2018]. Synthetic data is produced in a completely artificial manner, while semi-synthetic data combines real and computer-generated components. Both approximate real data and are used in their place to train deep learning methods or validate mathematical models [Nikolenko et al. 2019]. In the addressed context, synthetic datasets consist of virtual hands generated with 3D modeling software and often look artificial and unrealistic [Shilkrot et al. 2019]. Instead, semi-synthetic datasets are composed of real hands captured in constrained environments, such as a studio setting with green screens, professional lighting, and static cameras. Chroma-key software and traditional background subtraction techniques are used to remove the green background and extract the foreground objects [Gonzalez-Sosa et al. 2020; Lin and Martinez 2020]. The foregrounds are finally composed with many background scenes to build large-scale datasets. The images obtained do not exhibit various lighting conditions, and the shadows are almost removed from the studio lights with high luminous intensity. Furthermore, the object shadows in the new background are absent, creating a composite image deprived of part of the natural depth aspect. The composite image usually has an artifact aspect, in which the foreground is not well blended with the background, and there are significant chromatic discrepancies. In addition, the usage of background images that contain salient objects belonging to the same class as the foreground can cause semantic ambiguities. For example, if people occur in the background and their hands are marked as background class, training with this data could mislead the neural network [Li et al. 2020]. Training deep learning models on a source domain that differs significantly from the target domain could lead to poor results in real-life case evaluations. Intensive pre-processing, such as image-to-image translation techniques [Capece et al. 2019; Mueller et al. 2018] and image harmonization methods [Tsai et al. 2017], is usually used to bring realism to synthetic and semi-synthetic data. Additionally, domain or scene adaptation may be required, e.g., fine-tuning the network using custom scene images to improve the accuracy for specific scenarios [Lin and Martinez 2020].

Although various realistic datasets are publicly available and can overcome the limitations of artificial datasets, they often contain low-quality images with a pixelated effect and coarse segmentation masks [Tang et al. 2018]. Table 1 illustrates the careful and thorough classification we made of the most popular real datasets available, consisting of real-life RGB images in egocentric vision and segmentation masks of hands and arms. We found that some datasets contain only the hand label, although the arm is also shown in the images, such as EgoHands [Bambach et al. 2015] and EYTH [Urooj and Borji 2018]. Others consist of images showing both clothed and bare arms, but only the bare arms are labeled as foreground, while the clothed ones are marked as the background, i.e., THU-READ [Tang et al. 2018], KBH [Wang et al. 2019], and a relatively small part of TEgO [Lee and Kacorri 2019]. Instead, EGTEA Gaze+ [Li et al. 2018] provides several coarse and polygonal labels and some mislabeled regions, e.g., objects causing occlusion identified as hand. Additionally, there is a small number of erroneous classifications in EDSH [Li and Kitani 2013] and EgoGestureSeg [Gonzalez-Sosa et al. 2020]. The latter refers to a small subset of EgoGesture dataset [Zhang et al. 2018] manually annotated by Gonzalez *et al.* [Gonzalez-Sosa et al. 2020]. It contains images with indoor and outdoor scenarios, varying ambient light conditions, shadows, occlusions between the hands, and a challenging volume of motion blur.

We collected a large upper limb segmentation dataset that extends the existing real datasets and overcomes their limitations.

*Web3D and Deep Learning.* Recently, DL techniques have been applied to Web3D. H. Kim *et al.* [Kim and Won Lee 2020] developed a DL-based recognition system for 3D objects stored in X3D files. In particular, they trained a deep neural network using the geometric coordinates of 3D models and estimating their shapes. Another approach for 3D model analysis was proposed by W. Zhou *et al.* [Zhou et al. 2020], who trained two Siamese networks based on the VGG-16 model for shape retrieval and introduced an hybrid convolutional neural network to obtain the best view of 3D furniture. The main drawback of using DL for web applications is the requirement of large computational resources. For example, several DL frameworks require at least one GPU with CUDA enabled. In addition, DL algorithms are usually developed in a certain programming language and run on a certain operating system or platform. Recently, researchers have then spent much effort on DL-powered web solutions and obtained promising results, although there is still a gap between desktop and in-browser approaches [Ma et al. 2019]. In particular, several in-browser DL frameworks allow performing network inference and training directly in the browser, e.g., WebDNN [Hidaka et al. 2017] and TensorFlow.js [Smilkov et al. 2019]. An interesting approach using this technology was developed by N. Xie *et al.* [Xie et al. 2019]. In detail, they trained

**Table 1: Real datasets showing both hands and arms. They are categorized into vision mode (FPV: first-person view or egocentric; TPV: third-person view), skin tone, clothes arms, limbs labeled as foreground (B: bare; C: clothed; C*: clothed with some errors), scenarios, lighting conditions (A: variable ambient light; F: flashlight; Sh: strong shadows; S: strong sunlight), occlusions (HO: hand-to-object; HH: hand-to-hand), presence of motion blur, label quality (Some coarse: some labels are coarse and polygonal; Coarse: most or all labels are coarse; Err.: some wrong classifications; Acc.: accurate data), and image quality (L-: very low; L: low; M: medium; H: high).**

| Dataset | Mode | Skin | Clothed Arms | Labeled Limbs | Scenario | Light | Occlusion | Motion Blur | Label Quality | Image Quality |
|---|---|---|---|---|---|---|---|---|---|---|
| EgoHands [Bambach et al. 2015] | FPV, TPV | White | ✓ | Hand | Indoor, Outdoor | A | HO, HH | ✓ | Some coarse | H |
| EGTEA Gaze+ [Li et al. 2018] | FPV | White | – | Hand, forearm (B) | Indoor | A | HO, HH | ✓ | Some coarse, Err. | H |
| THU-READ [Tang et al. 2018] | FPV | White | ✓ | Hand, forearm (B) | Indoor, Outdoor | A, Sh | HO, HH | ✓ | Coarse | L- |
| KBH [Wang et al. 2019] | FPV | White | ✓ | Hand, forearm (B) | Indoor | A, F, Sh | – | ✓ | Some coarse | L |
| EYTH [Urooj and Borji 2018] | FPV, TPV | White | ✓ | Hand | Indoor, Outdoor | A | HO, HH | ✓ | Acc. | L |
| EgoGestureSeg [Gonzalez-Sosa et al. 2020] | FPV | White | ✓ | Hand, forearm (B, C) | Indoor, Outdoor | A, S, Sh | HH | ✓ | Acc., Err. | M |
| EDSH [Li and Kitani 2013] | FPV | White | – | Hand, forearm (B) | Indoor, Outdoor | A, S, Sh | HO | ✓ | Acc., Err. | M |
| TEgO [Lee and Kacorri 2019] | FPV | White, Black | ✓ | Hand, forearm (B, C*) | Indoor | A, F, Sh | HO | – | Acc., Err. | H |
| EgoCam (ours) | FPV | White | ✓ | Hand, forearm(B, C), elbow (C), part of the upper-arm (C) | Indoor, Outdoor | A, S, Sh | HH | ✓ | Acc. | H |

a conditional generative adversarial network to generate global illumination maps of 3D human organs and used TensorFlow.js for network prediction directly on the Web3D client. Those frameworks also provide a model converter, which converts and optimizes neural networks trained with well-known native DL frameworks to enable fast execution on web browsers. In addition, they support backend implementations for JavaScript APIs (such as WebGL and WebGPU) that can be used to access the GPU and speed up DL algorithms, such as in Web3D applications.

## 3 UPPER LIMB SEGMENTATION DATASET

To collect the upper limb segmentation dataset, we established strict requirements to select the best images from existing datasets or acquire new data. In particular, we focused on many interesting parameters, such as the vision mode (first-person or third-person view), the diversification of skin tone, scenarios, and lighting conditions, the presence of motion blur due to realistic user/camera movements in egocentric vision, inter-hand and hand-to-object occlusions. Moreover, we investigated the quality of the data: *i)* in the case of images, we considered the resolution and defined as a positive factor the absence of artifacts and compressions that cause a pixelated effect; *ii)* in the case of labels, we considered polygonal and coarse ground-truth masks and labeling errors as bad factors (for example, objects misclassified as foreground and arms partially labeled). Finally, we checked the presence of bare or clothed arms in the images. Since we are interested in both cases, both bare and clothed arms must be labeled as foreground.

As can be noted in Table 1 and Figures 2 and 3, EDSH [Li and Kitani 2013] and TEgO [Lee and Kacorri 2019] mostly met the main requirements and show different cases of lighting and skin tones. In particular, the first image of Figure 2 was taken from the EDSH subset that includes hand-to-object interactions in the kitchen, and the other three show variable ambient light, shadows, and strong sunlight conditions. Instead, Figure 3 highlighted the diversity of skin tone, hand-to-object occlusions, and different lighting cases of the TEgO dataset. The first image comes from the in-the-wild subset with flashlight illumination, the second is from the in-the-wild subset with no indoor lighting, and the last two come from the in-the-vanilla subset. Hence, we used their best data as a part of the upper limb segmentation dataset, discarding incorrect or partially labeled data. The correctness of the labels was manually verified by carefully inspecting the data and overlapping each RGB image with the corresponding ground-truth mask.

The second part of the upper limb segmentation dataset consists of our dataset, called EgoCam. It was built using two cameras in egocentric vision: the first device had an extra-wide field of view and $1,280 \times 720$ resolution, while the second device had a standard field of view and higher resolution, i.e., $1,080 \times 1,920$. Both cameras were employed to record RGB videos at 30 fps. Four subjects (two male and two female) were shot moving freely in indoor or outdoor scenes, with a wide variety of lighting conditions, backgrounds, and clothes (see Table 1 for details). Moreover, the limbs and hands were framed in different poses, and the space occupied in the images was variable. For example, the hand and arm can be near or far from the

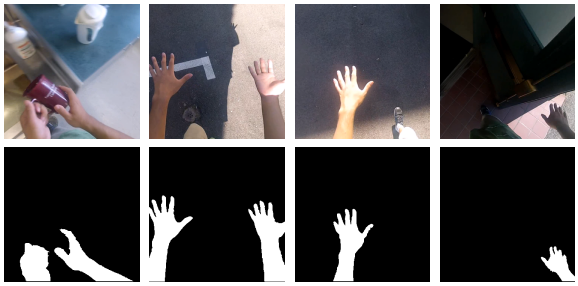

**Figure 2: Some example of images and ground-truth annotations of the EDSH dataset [Li and Kitani 2013].**

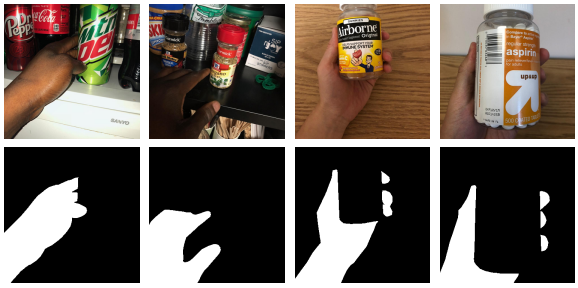

**Figure 3: Some images and ground-truth annotations of the TEgO dataset [Lee and Kacorri 2019].**

**Figure 4: Some images and well-annotated masks of our EgoCam dataset.**

camera, enter or leave the field of view from different directions, almost fill the image or occupy a small region of it, as shown in Figure 4. In detail, the arm, elbow, and part of the upper arm are captured in the first two images. The third photo is an example of inter-hand occlusion, while the fourth and the fifth show some blur near the fingers caused by the movement. The last three panels display hands and arms with different lights and clothes.

EgoCam has two main advantages. Firstly, it increases the number of well-annotated images in the upper limb segmentation dataset. In particular, all data was manually labeled. Secondly, it contains video frames showing cases that are poor or absent in all analyzed datasets. In details, the elbow and upper arm were captured, since those limb sections can be easily framed using egocentric cameras with a wider field of view, such as the GoPro and ZED Mini. In addition, we introduced arms with clothes that differ in color, fabric, shape, and length, providing further information variety.

Since the collected data had different orientations (portrait or landscape) and aspect ratio, we performed a square crop designing an automatic procedure that moves the cropping area with a fixed step for each dataset and saves the square images and labels. In particular, the procedure defined a square cutting bounding box with dimensions equal to $\min(h, w)$, where $h$ and $w$ were the height and width of the original image. The bounding box was then automatically moved along the largest dimension of the original image, i.e., vertically for portrait images and horizontally for the landscape ones. A shift step of 80 was used for the EgoCam dataset, while it was set to 30 for the TEgO and EDSH datasets. A different shift step was adopted since the limb occupies a smaller portion inside the frame in the case of the EgoCam dataset. Therefore, a large number of crops would lead to many images similar to each other and where the limb is not often captured. Finally, we inspected all saved square images and deleted any useless clippings, e.g., with no hands, arms, or distinguishable parts of them. The main advantage of that procedure is that a further and greater variety of limb positions is provided, as hands and arms can also be partially visible and located in different image portions. Thus, our final overall dataset on upper limb segmentation consisted of $46,021$ well-annotated frames: $1,165$ from EDSH, $31,999$ from TEgO, and $12,857$ from EgoCam. It was split into $43,837$ images for training and $2,184$ for testing. The training set contains images taken from all three datasets, while the test set includes only a subset of $2,079$ TEgO and 709 EgoCam data that differ from the training set. This choice is due to the limited number of EDSH images, i.e., $1,165$ frames showing only a single subject's upper limbs. Finally, all data was reshaped to the spatial size of $360 \times 360$ to accelerate the network training.

## 4 THE PROPOSED APPROACH

Our goal was to accurately segment human upper limbs in egocentric unconstrained real-life environments. We propose an end-to-end deep learning approach, which is robust to inter-hand and hand-object occlusions, various lighting conditions, different limb position in the frame and skin tone, indoor and outdoor scenarios, dynamic user/camera movements. In particular, we focused on the DeepLabv3+ model [Chen et al. 2018] since it is the state-of-the-art networks for semantic segmentation and achieved notable results on many benchmark datasets. It is characterized by an encoder-decoder architecture. The encoder extracts semantic information

and reduces the size of the feature maps, whereas the decoder recovers spatial and detailed object boundary information. The DeepLabv3+ encoder is based on the DeepLabv3 model [Chen et al. 2017b], which combines the advantages of the atrous (or dilated) convolutions and atrous spatial pyramid pooling (ASPP) introduced by the previous DeepLab architectures [Chen et al. 2014, 2017a]. In particular, the atrous convolution is a modified version of the convolution operation, which upsamples convolution filters with holes in order to enlarge their field of view and take the context more into account, but without increasing the computational costs and number of parameters. The ASPP module consists of several layers placed in parallel: one $1 \times 1$ convolution, three atrous convolutions with multiple atrous rates, and an image pooling layer. The main benefit of ASPP is the ability to capture the multi-scale context effectively and improve segmentation. The overall structure of the DeepLabv3+ encoder consists of a backbone network and an ASPP module followed by a $1 \times 1$ convolutional layer. The backbone is usually a deep convolutional neural network that employs atrous convolutions. The decoder exploits the information learned from both the encoder and the backbone. Indeed, the low-level features of the backbone network are passed through a $1 \times 1$ convolution and then concatenated with the encoder output, which is first upsampled using bilinear interpolation. Finally, a $3 \times 3$ convolution and a further bilinear upsampling are applied.

We trained several networks based on the DeepLabv3+ architecture using the upper limb segmentation dataset and conducted many experiments to find a robust and efficient configuration. Hence, we considered various backbone networks based on the ResNet [He et al. 2016] and Xception [Chollet 2017] models adapted to semantic segmentation. In particular, the first two networks were based on the beta variant of ResNet, a modified version of the original ResNet in which the first $7 \times 7$ convolution was replaced with three $3 \times 3$ convolutions. The backbone networks had 50 and 101 layers and were named ResNet-50-beta and ResNet-101-beta, respectively. The other two backbone networks were based on the modified aligned Xception network proposed by Chen *et al.* [Chen et al. 2018], which is an effective network that showed promising results and a reduction in computation time. In detail, depthwise separable convolution replaced max pooling operations of the original Xception network [Chollet 2017], and further batch-normalization and ReLU were added. We adopted the Xception backbone networks with 41 and 65 layers.

All backbone networks were trained using the cross-entropy loss function and the stochastic gradient descent optimization algorithm with momentum set to 0.9 and using polynomial learning rate policy with 0.9 power value, following the training protocol suggested in [Chen et al. 2017b]. Training a deep model from scratch requires a copious amount of data and resources in terms of memory, computation, and time. Although GPU acceleration is widely used, it is usually recommended to start from a pre-trained model on a huge dataset (e.g., millions of data) and apply transfer learning or fine-tuning techniques. Therefore, we used publicly available pre-trained weights[2]. We found models pre-trained on the ImageNet [Russakovsky et al. 2015] and MS-COCO [Lin et al. 2014] datasets in the case of Xception-65, while only ImageNet pre-trained weights for the ResNet-50-beta, ResNet-101-beta, and Xception-41 based networks. ImageNet is a generic and huge dataset, often used for classification and object detection tasks, while MS-COCO is smaller than ImageNet and used for classification, detection, and segmentation. For this reason, a segmentation network pre-trained on MS-COCO may benefit more from the learned features than using only ImageNet pre-training [Chen et al. 2018], and a shorter training may be sufficient. Then, Deeplabv3+ built on top of the ResNet and smaller Xception networks were trained for 300K iterations, with batch size set to 12, and 0.001 base learning rate. The training of the other model lasted 90K iterations until convergence, setting the batch size to 8 and the base learning rate to 0.0001. All models were trained using one Nvidia Titan Xp GPU with 12GB memory, and data augmentation with random left/right flip was applied during training to avoid model overfitting.

[2]The pre-trained weights used are publicly available on the DeepLab project page: https://github.com/tensorflow/models/tree/master/research/deeplab

## 5 RESULTS

We evaluated the effectiveness of our approach and the robustness against several scenarios and conditions, comparing the trained models. We performed both quantitative and qualitative extensive tests. In particular, we tested all models on our test set and EgoGestureSeg dataset to assess their performance and generalization level. In the case of quantitative tests, we adopted conventional metrics for the segmentation task [Minaee et al. 2021], such as Accuracy (Acc), Intersection over Union (IoU), and mean F1 score (mF1) for each category (upper limb and background classes). Since Acc and IoU may not be reliable for the overall assessment of the test set (e.g., due to unbalanced classes), the average over the number of categories is usually calculated, obtaining mAcc and mIoU, respectively. [Lateef and Ruichek 2019]. For graphic purposes, the DeepLabv3+ backbone networks were indicated by abbreviations in the following tables and figures (X-41: Xception-41; X-65: Xception-65; R-50-b: ResNet-50-beta; R-101-b: ResNet-101-beta).

### 5.1 Results on Upper Limb Segmentation Test Set

Firstly, we tested all models on the test subset of our upper limb segmentation dataset. The network with ResNet-101-beta backbone performed best on most of the metrics considered, as illustrated in Table 2. In the case of the mAcc metric for each class, such a network achieved the second-best result. In particular, it differs by 0.0006 for the background class and a negligible amount of 0.0001 for the upper limb class compared to the best case. However, the obtained values show excellent results for all trained networks. The worst model often diverges by limited quantities and with a difference that does not exceed 1-2%. On the other hand, it is not easy to identify the best network observing the qualitative results on the test set. Some examples are shown in Figure 5. The network with ResNet-101-beta backbone provides the most accurate masks, although there are some pixel classification errors between the fingers in the second and fourth columns. In these two cases, the most accurate is the Xception-65 based network. Better precision, especially on the edges of hands, can be noted. However, this network presents some wrong pixel classifications, as shown in the first and fifth columns of Figure 5. For this reason, we conducted other experiments on the

EgoGestureSeg (Section 5.2) to highlight the differences between the performance of our trained models.

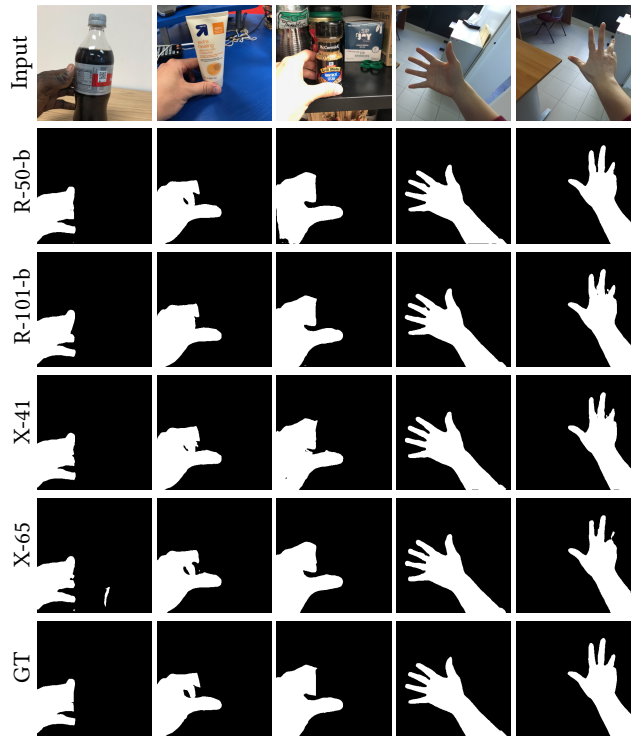


**Figure 5: Some qualitative results on test images from the upper limb segmentation dataset. The output of each network and the ground-truth segmentation masks are reported.**

## 5.2 Results on EgoGestureSeg

To further compare all models and evaluate the level of generalization, we tested them on the challenging EgoGestureSeg dataset (see Section 2 and Table 1 for details), providing both quantitative and qualitative assessments. Table 3 illustrates the metric values obtained. In this case, the model with the Xception-65 backbone showed better performance for almost all metrics. Interestingly, the gap between the best and the worst models is significant, e.g., for the limb class, there is a variation of about 14% on the mean Accuracy and 12% on the mean Intersection over Union and mean F1 score. That is most evident in the qualitative results in Figure 6. The input images in the first row show various lighting conditions and scenarios. In particular, the upper limbs are located only in a part of the image close to the frame edges (first and last panel), are strongly illuminated (second panel) or backlit and dark (fourth panel). In addition, a cluttered background in the third image and a very blurry hand in the last one are shown. As can be noted, ResNet-based models often failed to fully locate the limb, especially on the last two images. A slight improvement occurred with the Xception-41 based network. Instead, the network that uses the Xception-65 backbone found the limb in all images with only minor errors. It is noteworthy that Xception-61 based network is the only model that did not make classification errors related to the background of the last image. In our opinion, the results obtained with the EgoGestureSeg dataset prove that using initial weights pre-trained with both ImageNet and MS-COCO can help the network have a broader view and be able to generalize better. Furthermore, the conducted tests show that the Xception-based architecture is more robust to challenging situations than the ResNet-based one.

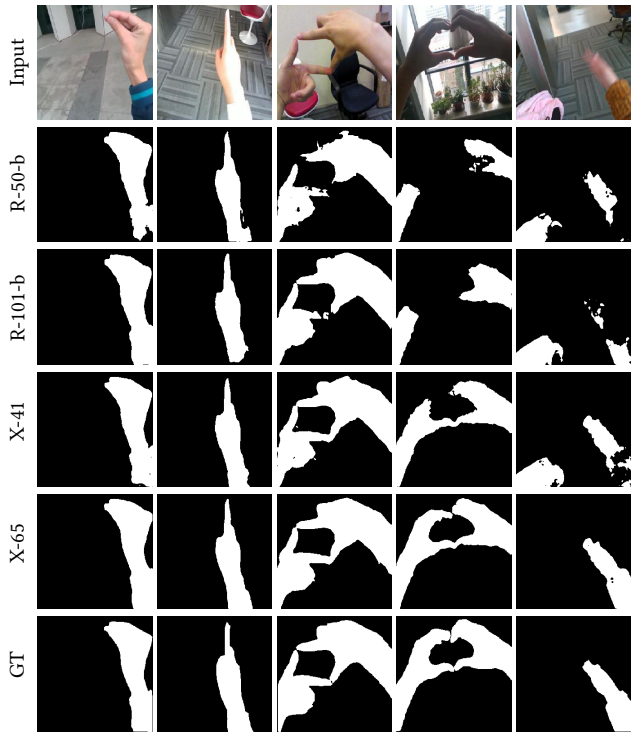


**Figure 6: Some qualitative results testing all models on the EgoGestureSeg dataset. The predictions of each model (from the second to the fifth row) and the ground-truth labels (last row) are shown.**

## 5.3 Inference Time

We made a further comparison between the models considering the inference time. In particular, we measured the average time required to segment RGB images with spatial dimensions equal to the network input size, i.e., $360 \times 360$, as shown in Table 4. All networks were tested on the same computer, equipped with one Nvidia Titan Xp GPU with 12GB memory. As expected, the deeper models based on Xception-65 and ResNet-101-beta were slower than models with fewer layers. Although the mean computation time of the deeper models was the same, the network based on the Xception-65 backbone is lighter than the one based on ResNet-101-beta. Furthermore, the average inference times of all models prove more than real-time performance.

## 6 CONCLUSION

In the proposed work, we focused on upper limb segmentation in egocentric vision providing an effective end-to-end deep learning

**Table 2: Quantitative comparison using our upper limb segmentation test set. The metric values in percentages related to the overall test set and for each class are reported. The best values for each metric are highlighted in bold.**

| Models | Overall | | | Limb | | | Background | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAcc | mIoU | mF1 | Acc | IoU | mF1 | Acc | IoU | mF1 |
| DeepLabv3+ (X-41) | 99.14 | 98.00 | 97.92 | **98.57** | 96.47 | 97.17 | 99.71 | 99.53 | 98.67 |
| DeepLabv3+ (X-65) | 98.84 | 97.69 | 97.35 | 97.96 | 95.93 | 96.31 | 99.72 | 99.46 | 98.39 |
| DeepLabv3+ (R-50-b) | 98.95 | 98.23 | 98.54 | 98.06 | 96.88 | 98.05 | **99.84** | **99.59** | 99.02 |
| DeepLabv3+ (R-101-b) | **99.17** | **98.26** | **98.63** | 98.56 | **96.94** | **98.19** | 99.78 | **99.59** | **99.07** |

**Table 3: Metric values in percentages computed testing all models on the EgoGestureSeg dataset. The metrics related to the overall set and for each class are reported. The best values are highlighted in bold.**

| Models | Overall | | | Limb | | | Background | | |
|---|---|---|---|---|---|---|---|---|---|
| | mAcc | mIoU | mF1 | Acc | IoU | mF1 | Acc | IoU | mF1 |
| DeepLabv3+ (X-41) | 95.02 | 92.64 | 87.84 | 90.64 | 87.29 | 82.83 | 99.40 | 97.98 | 92.80 |
| DeepLabv3+ (X-65) | **96.05** | **93.38** | **90.42** | **92.83** | **88.59** | **86.71** | 99.25 | **98.16** | **94.13** |
| DeepLabv3+ (R-50-b) | 89.16 | 86.50 | 82.44 | 78.77 | 76.61 | 74.35 | **99.56** | 96.39 | 90.39 |
| DeepLabv3+ (R-101-b) | 92.30 | 89.85 | 85.88 | 85.10 | 82.44 | 79.47 | 99.50 | 97.25 | 92.24 |

**Table 4: Average inference time computed using one Nvidia Titan Xp GPU with 12GB memory and testing RGB images with a spatial dimension equal to the network input size, i.e., $360 \times 360$. The model size for each neural network is also reported.**

| Models | Inference time (s) | Model size (MB) |
|---|---|---|
| DeepLabv3+ (X-41) | 0.017 | 108 |
| DeepLabv3+ (X-65) | 0.02 | 158 |
| DeepLabv3+ (R-50-b) | 0.016 | 102 |
| DeepLabv3+ (R-101-b) | 0.02 | 175 |

solution in unconstrained real-life environments. Since existing approaches often need a large amount of well-annotated data to learn from, many datasets were developed. Most of them employed a large quantity of synthetic and semi-synthetic images, which are usually labeled with less effort and cost than real data but differ considerably from the target realistic domain in terms of chromaticity, lighting conditions, foreground/background blending, and overall appearance. That could lead to poor model performance in real-world cases, and domain/scene adaptation techniques can be required. On the other hand, obtaining accurate segmentation masks from real photos and video frames is laborious and sometimes impractical. Although various realistic datasets are publicly available, they often contain low-quality images and coarse segmentation masks. Furthermore, existing approaches based on real data were limited to the segmentation of hands up to the wrist or bare arm. To overcome those limitations, we trained several segmentation deep neural networks based on the state-of-the-art DeepLabv3+ model. Moreover, we collected a large well-annotated upper limb segmentation dataset. It contains about 46 thousand images and shows a wide range of real-life scenarios, lighting conditions, skin tone, clothes, and occlusions. The collected dataset consists of carefully selected data from the TEgO and EDSH datasets, which met the main requirements we looked for, and our manually labeled EgoCam dataset. The latter was built using two cameras in egocentric vision, recording videos with various female and male subjects in a wide variety of situations. Finally, we extensively tested and compared the trained networks to find the best model configuration, assess their generalization level, and prove the robustness against several scenarios, achieving remarkable results and more than real-time performances. The main advantage of the proposed approach regards the possibility of accurately segment both bare and clothed human upper limb, also in case of inter-hand and hand-object occlusions, variable lighting conditions, skin colors, indoor and outdoor real-life environments. It can be particularly useful when egocentric cameras with a wide field of view are employed since they can capture the whole upper limb and not only the user's hand. To the best of our knowledge, the proposed work is the first to evaluate the effectiveness of deep learning approach in such unconstrained real-world scenarios and collecting a large comprehensive dataset with real well-annotated images. It could allow locating the user's limb interactively and also obtaining pixel-precise information. That may increase the user's sense of presence and body ownership, for example, in Web3D VEs, providing a general-purpose solution that can be executed on any device using a web browser and without a specific hardware configuration or software installation. In the future, we plan to conduct a user study to assess the usefulness of our approach in web-based immersive applications and perform a comparison with state-of-the-art methods for egocentric hand and arm segmentation.

## REFERENCES

Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. 2015. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE International Conference on Computer Vision*. 1949–1957.

Andrea Bandini and José Zariffa. 2020. Analysis of the hands in egocentric vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* (2020).

Alejandro Betancourt, Pietro Morerio, Emilia Barakova, Lucio Marcenaro, Matthias Rauterberg, and Carlo Regazzoni. 2017. Left/right hand segmentation in egocentric videos. *Computer Vision and Image Understanding* 154 (2017), 73–81.

Alejandro Betancourt, Pietro Morerio, Carlo S Regazzoni, and Matthias Rauterberg. 2015. The evolution of first person vision methods: A survey. *IEEE Transactions on*

*Circuits and Systems for Video Technology* 25, 5 (2015), 744–760.

Giuseppe Caggianese, Luigi Gallo, and Pietro Neroni. 2015. Design and preliminary evaluation of free-hand travel techniques for wearable immersive virtual reality systems with egocentric sensing. In *International Conference on Augmented and Virtual Reality*. Springer, 399–408.

Giuseppe Caggianese, Luigi Gallo, and Pietro Neroni. 2016. An investigation of leap motion based 3D manipulation techniques for use in egocentric viewpoint. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, 318–330.

Minjie Cai, Kris M Kitani, and Yoichi Sato. 2017. An ego-vision system for hand grasp analysis. *IEEE Transactions on Human-Machine Systems* 47, 4 (2017), 524–535.

Minjie Cai, Feng Lu, and Yoichi Sato. 2020. Generalizing Hand Segmentation in Egocentric Videos With Uncertainty-Guided Model Adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14380–14389. https://doi.org/10.1109/CVPR42600.2020.01440

Nicola Capece, Francesco Banterle, Paolo Cignoni, Fabio Ganovelli, Roberto Scopigno, and Ugo Erra. 2019. Deepflash: Turning a flash selfie into a studio portrait. *Signal Processing: Image Communication* 77 (2019), 28–39.

Nicola Capece, Ugo Erra, Monica Gruosso, and Marco Anastasio. 2020. Archaeo Puzzle: An Educational Game Using Natural User Interface for Historical Artifacts. (2020).

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2014. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062* (2014).

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 2017a. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* 40, 4 (2017), 834–848.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017b. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*.

Luca Chittaro and Milena Serra. 2004. A brief introduction to Web3D technologies in education: Motivations, issues, opportunities. In *Proceedings of LET-WEB3D 2004, the First International Workshop on Web3D Technologies in Learning, Education and Training*. Citeseer, 3–7.

François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

Marcelo de Paiva Guimarães, Diego Roberto Colombo Dias, José Hamilton Mota, Bruno Barberi Gnecco, Vinícius Humberto Serapilha Durelli, and Luis Carlos Trevelin. 2018. Immersive and interactive virtual reality applications based on 3D web browsers. *Multimedia Tools and Applications* 77, 1 (2018), 347–361.

Ester Gonzalez-Sosa, Pablo Perez, Ruben Tolosana, Redouane Kachach, and Alvaro Villegas. 2020. Enhanced Self-Perception in Mixed Reality: Egocentric Arm Segmentation and Database With Automatic Labeling. *IEEE Access* 8 (2020), 146887–146900.

Monica Gruosso, Nicola Capece, and Ugo Erra. 2021. Human segmentation in surveillance video with deep learning. *Multimedia Tools and Applications* 80, 1 (2021), 1175–1199.

Monica Gruosso, Nicola Capece, Ugo Erra, and Francesco Angiolillo. 2020. A Preliminary Investigation into a Deep Learning Implementation for Hand Tracking on Mobile Devices. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 380–385.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Masatoshi Hidaka, Yuichiro Kikura, Yoshitaka Ushiku, and Tatsuya Harada. 2017. Webdnn: Fastest dnn execution framework on web browser. In *Proceedings of the 25th ACM international conference on Multimedia*. 1213–1216.

Tin Hok, Chan Park, and Kwan-Hee Yoo. 2020. System Architecture for Supporting Multiple Live Actors in Web3D Virtual Conference. In *The 25th International Conference on 3D Web Technology*. 1–4.

Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments* 21, 4 (2012), 373–387.

Ha-Seong Kim and Myeong Won Lee. 2020. 3D Object Recognition Using X3D and Deep Learning. In *The 25th International Conference on 3D Web Technology*. 1–8.

Fahad Lateef and Yassine Ruichek. 2019. Survey on semantic segmentation using deep learning techniques. *Neurocomputing* 338 (2019), 321–348.

Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Cheng Li and Kris M Kitani. 2013. Pixel-level hand detection in ego-centric videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3570–3577.

Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. 2020. End-to-end Animal Image Matting. *arXiv preprint arXiv:2010.16188* (2020).

Yinlin Li, Lihao Jia, Zidong Wang, Yang Qian, and Hong Qiao. 2019. Un-supervised and semi-supervised hand segmentation in egocentric images with noisy label learning. *Neurocomputing* 334 (2019), 11–24. https://doi.org/10.1016/j.neucom.2018.12.010

Yin Li, Miao Liu, and James M Rehg. 2018. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 619–635.

Jirapat Likitlersuang, Elizabeth R Sumitro, Tianshi Cao, Ryan J Visée, Sukhvinder Kalsi-Ryan, and José Zariffa. 2019. Egocentric video: a new tool for capturing hand use of individuals with spinal cord injury at home. *Journal of neuroengineering and rehabilitation* 16, 1 (2019), 1–11.

Fanqing Lin and Tony Martinez. 2020. Ego2Hands: A Dataset for Egocentric Two-hand Segmentation and Detection. *arXiv preprint arXiv:2011.07252* (2020).

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

Yun Ma, Dongwei Xiang, Shuyu Zheng, Deyu Tian, and Xuanzhe Liu. 2019. Moving deep learning into web browser: How far can we go?. In *The World Wide Web Conference*. 1234–1244.

Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. 2021. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 49–59.

Supreeth Narasimhaswamy, Zhengwei Wei, Yang Wang, Justin Zhang, and Minh Hoai. 2019. Contextual attention for hand detection in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9567–9576.

Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, and Francisco Florez-Revuelta. 2018. Recognition of Activities of Daily Living from Egocentric Videos Using Hands Detected by a Deep Convolutional Network. In *Image Analysis and Recognition*, Aurélio Campilho, Fakhri Karray, and Bart ter Haar Romeny (Eds.). Springer International Publishing, Cham, 390–398.

Sergey I Nikolenko et al. 2019. Synthetic data for deep learning. *arXiv preprint arXiv:1909.11512* 3 (2019).

Sai Krishna Pathi, Andrey Kiselev, Annica Kristoffersson, Dirk Repsilber, and Amy Loutfi. 2019. A novel method for estimating distances from a robot to humans using egocentric RGB camera. *Sensors* 19, 14 (2019), 3142.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

Roy Shilkrot, Supreeth Narasimhaswamy, Saif Vazir, and Minh Hoai. 2019. WorkingHands: A Hand-Tool Assembly Dataset for Image Segmentation and Activity Mining.. In *BMVC*. 258.

Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, et al. 2019. Tensorflow. js: Machine learning for the web and beyond. *arXiv preprint arXiv:1901.05350* (2019).

Yansong Tang, Zian Wang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. 2018. Multi-stream deep neural networks for rgb-d egocentric action recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 3001–3015.

Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3789–3797.

Aisha Urooj and Ali Borji. 2018. Analysis of hand segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4710–4719.

Wei Wang, Kaicheng Yu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. 2019. Recurrent U-Net for resource-constrained segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2142–2151.

Ning Xie, Yifan Lu, and Chang Liu. 2019. Web3D client-enhanced global illumination via GAN for health visualization. *IEEE Access* 8 (2019), 13270–13281.

Juan Zamora-Mora and Mario Chacón-Rivas. 2019. Real-Time Hand Detection using Convolutional Neural Networks for Costa Rican Sign Language Recognition. In *2019 International Conference on Inclusive Technologies and Education (CONTIE)*. IEEE, 180–1806.

Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. 2018. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia* 20, 5 (2018), 1038–1050.

Wen Zhou, Jinyuan Jia, Chengxi Huang, and Yongqing Cheng. 2020. Web3D learning framework for 3D shape retrieval based on hybrid convolutional neural networks. *Tsinghua Science and Technology* 25, 1 (2020), 93–102. https://doi.org/10.26599/TST.2018.9010113

Christian Zimmermann and Thomas Brox. 2017. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*. 4903–4911.