IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 10, 2020, accepted July 27, 2020, date of publication July 30, 2020, date of current version August 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3013016

Enhanced Self-Perception in Mixed Reality: Egocentric Arm Segmentation and Database With Automatic Labeling

ESTER GONZALEZ-SOSA[1], PABLO PÉREZ[1], RUBEN TOLOSANA[2], REDOUANE KACHACH[1], AND ALVARO VILLEGAS[1]

[1]Nokia Bell Labs, 28050 Madrid, Spain
[2]Departamento de Tecnología Electrónica y de las Comunicaciones, Universidad Autónoma de Madrid, 28049 Madrid, Spain

Corresponding author: Ester Gonzalez-Sosa (ester.gonzalez@nokia-bell-labs.com)

This work was supported in part by the Torres Quevedo Fund "PTQ-17-09374" from the Ministerio de Ciencia, Innovación y Universidades.

ABSTRACT In this study, we focus on the egocentric segmentation of arms to improve self-perception in Augmented Virtuality (AV). The main contributions of this work are: *i*) a comprehensive survey of segmentation algorithms for AV; *ii*) an *Egocentric Arm Segmentation Dataset (EgoArm)*, composed of more than 10, 000 images, demographically inclusive (variations of skin color, and gender), and open for research purposes. We also provide all details required for the automated generation of groundtruth and semi-synthetic images; *iii*) the proposal of a deep learning network to segment arms in AV; *iv*) a detailed quantitative and qualitative evaluation to showcase the usefulness of the deep network and EgoArm dataset, reporting results on different real egocentric hand datasets, including GTEA Gaze+, EDSH, EgoHands, Ego Youtube Hands, THU-Read, TEgO, FPAB, and Ego Gesture, which allow for direct comparisons with existing approaches using color or depth. Results confirm the suitability of the EgoArm dataset for this task, achieving improvements up to 40% with respect to the baseline network, depending on the particular dataset. Results also suggest that, while approaches based on color or depth can work under controlled conditions (lack of occlusion, uniform lighting, only objects of interest in the near range, controlled background, etc.), deep learning is more robust in real AV applications.

INDEX TERMS Egocentric arm segmentation, mixed reality, augmented virtuality, self-perception, arm segmentation, automatic labeling, EgoArm dataset, demographically inclusive.

I. INTRODUCTION

Most computer vision applications are traditionally focused on third-person view (TPV) actions that happen while interacting directly or indirectly with a camera [6]. With the advent of new wearable devices such as GoPro, Microsoft SenseCam, or even some Head Mounted Displays (HMD) used in immersive applications, research on first-person view (FPV) or egocentric vision is attracting some attention [7]. Main research lines in egocentric vision can be categorized into:

- **Localize egocentric objects** usually knowing hand position and recognizing which objects are in contact with them. Typical tasks here are recognition [53], detection [35], segmentation [59], tracking, and prediction [69], among others.
- **Recognize the activities** performed by humans through the analysis of the relationship between objects and hands (action recognition) [15], [20] or hand gesture recognition for Virtual Reality (VR) and Augmented Reality (AR) [10], [14], [50], hand poses [39], etc.
- **Visual lifelogging,** which consists on capturing daily experiences [8]. Video summarization of people lives is also a related area, which could be used for detecting novel or anomalous events. This research line is of special relevance for people with memory loss [17].

In this study, we explore egocentric arm segmentation as an essential requirement for enhanced self-perception in Mixed Reality (MR) (see Fig. 1). One of the main problems of immersive environments (IE)[1] is the so-called

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

[1]Immersive Environment covers Virtual Reality environments (computer generated or 360° video), and also Mixed Reality environments, combining IE with the reality surrounding the user (hereinafter local reality).

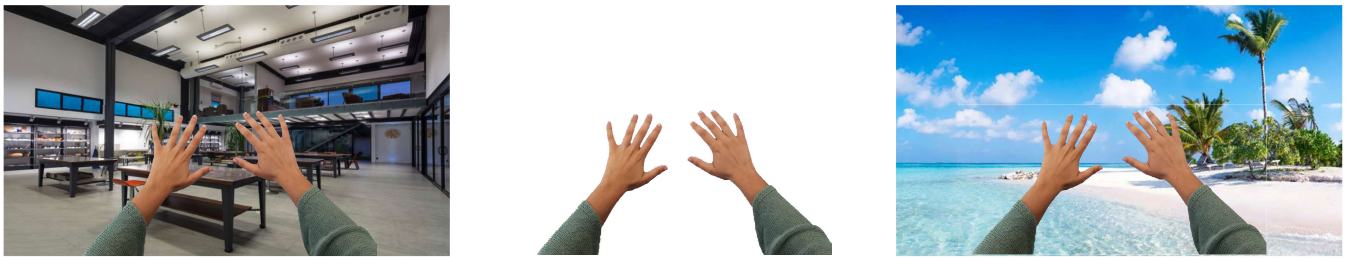

FIGURE 1. We propose semantic segmentation networks to segment human body parts (in this study the arms) to enhance self-perception in AV. Left: local reality; center: segmented arms; and right: AV with egocentric arms.

presence factor or *sense of presence*: the subjective experience of being in a remote location without moving from the physical place. According to Lee [34], the presence concept can be divided into three components: physical, social, and self-presence. In particular, self-presence involves *experiencing the representation of one's own genuine self, physically or psychologically manifested, inside a virtual environment*.

First attempts towards self-perception on IE are based on avatars, which are virtual representations of the user mimicking his/her movements [54]. Current research lines further study avatar representations [2], [65], their effect on the user [3], [58], and their interaction with the IE [3], [21], [25], [30], [55].

Considering MR in particular, there is a different way of reaching self-perception. As stated by Milgram and Kishino [46] and Regenbrecht *et al.* [52], Augmented Virtuality (AV) is a MR subcategory of the virtuality continuum that aims to merge the reality surrounding the user (hereinafter local reality) with an IE. This means, instead of seeing an avatar of the user's body tracking his movements, the user is presented with his real body immersed in the IE. The merge of a real and virtual world can be achieved with the video see-through capabilities of the newest HMD devices such as HTC VIVE Pro, Varjo XR-1 or just by attaching a local camera to the HMD. Hence, human body parts (such as hands, arms, lower body, etc.) or local objects (such as keyboards, smartphones, coffee cups, etc.) can be segmented from the see-through video and merged into the IE. Depending on the objects segmented, AV could be used to: *i*) increase self-presence and/or awareness of other people to prevent isolation, or *ii*) facilitate interaction with local objects [44].

Main segmentation approaches proposed in the literature for AV have been based on color or depth. However, they still show some limitations such as very complex physical setups [44], limited field of view [32], or poor depth estimation that prevent AV from reaching its full potential. To overcome these limitations, we explore Semantic Segmentation algorithms (hereinafter Sem-Seg) proposed in the literature, based on deep learning (DL) to segment egocentric **arms**. Our main motivation to focus on the whole arms and not just the hands is to study this problem under real-life conditions. Indeed, arms and not just hands are easily visible when wearing a HMD (see Fig.1). Moreover, we also hypothesize that seeing your whole arms and not just your hands, may have a positive impact on the self-presence factor of the experience. Aside from the segmentation challenges pertaining to egocentric vision, the reader should notice that arms contain additional variability factors such as clothes or skin color that need to be considered. The proposed work is a continuation of [27] that we have significantly extended with the following contributions:

- a **comprehensive discussion** on segmentation methods for AV, conceptually categorizing them by color, depth and other approaches.
- an **EgoArm** dataset, composed of more than 10,000 semi-synthetic images, which is demographically inclusive and publicly available for research purposes.[2] In addition, we describe the procedure carried out to automatically generate the groundtruth mask.
- a proposal based on **deep segmentation networks** to segment egocentric arms. To the best of our knowledge, we are the first ones to consider deep networks for AV applications and the first to consider the whole arms and not just the hands.
- a **thorough and in-depth evaluation** of our method to a wide number of real egocentric datasets existing in the literature: GTEA Gaze+ [41], EDSH [36], EgoHands [5], Ego YoutubeHands [63], THU-READ [61], TEgO [33], FPAB [24], and Ego Gesture [70]. We also contribute with a segmentation groundtruth of representative subset from two RGB-D datasets: Ego Gesture (277) and THU-READ (203), which will also be made available for research purposes.
- a **comparison with former segmentation approaches for AV,** based on color or depth, highlighting their pros and cons.

The rest of this article is structured as follows: Section II covers related works regarding AV, with an emphasis on the different up-to-date algorithms proposed to segment local reality objects. Section III describes the *EgoArm Dataset* and the whole procedure to generate semi-synthetic images while automatically obtaining the segmentation groundtruth. Section IV presents the Sem-Seg algorithms considered to

[2] https://cloud.proinnovation.es/index.php/s/tekqtneGXgrUgFD

TABLE 1. Related works on egocentric human body parts or object segmentation for AV. They are categorized into color, depth or other approaches. CRF stands for conditional random fields. Notice that [59], [63] are focused on egocentric hand segmentation but are not designed for AV.

Work	Category	Framework	Augmented Element	Segmentation Method	Observations	User Experience
Metzger (1993) [45]	Color	Virtual Research Flight Helmet Toshiba IK-M40A cameras	Hands	Blue chroma-key	Illumination dependent	No
McGill *et al.* (1993) [44]		Oculus Rift DK1 1.8mm M12 wide-angle board lens	Hands Keyboard	Green chroma-key	Illumination dependent	Yes (108 subjects)
Bruder *et al.* (2009) [9]		eMagin 3DVisor Z800 2 USB cameras	Hands Body	Skin detection Floor subtraction	Brightness adjustment	Yes (7 subjects)
Gunther *et al.* (2015) [29]		Oculus Rift GoPro Hero3+	Hands	Color based on HSV	UX allows adjust parameters	No
Zhu *et al.* (2016) [73]		HTC Vive Kinect	Whole Body	Green Chroma	–	No
Perez *et al.* (2018) [49]		Samsung Gear HMD Samsung S8	Hands Food	Color based on Cr from YCrCb	Scalable to other objects	Yes (66 subjects)
Tecchia *et al.* (2014) [62]	Depth	Oculus DK1 3D camera	Hands Body	Depth	Virtual objects manipulation through fingertip tracking	No
Nahon *et al.* (2015) [48]		Oculus Rift DK1 Kinect v2	Own Body Local objects Other People	Depth	Kinect placed opposite to the user Point cloud not very dense	No
Lee *et al.* (2016) [32]		Oculus Rift DK2 Soft Kinect DS325	User's body	Depth	Head shaking for transition	Yes (14 subjects)
Alaee *et al.* (2018) [1]		Oculus Rift DK2 Intel Real Sense	Hands Smartphone	Depth in $10 - 40$ cm	Scalable to other objects	Yes (25 subjects)
Rauter *et al.* (2019) [51]		HTC Vive Pro	Near Real World Objects	Depth + post processing	Flexible depth range	No
Fiore et Interrante (2012) [22]	Other	NVIS SX60 2 Logitech C615	Hands	Foreground and Background Histograms	–	No
Serra *et al.* (2013) [59]		–	Hands	Random Forest Light + Time + Space Consistency	Gesture Recognition	No
Desai *et al.* (2017) [18]		Oculus Rift DK2 Leap Motion	Smartphone	Smartphone edge detection + statistical classifier + App sending screenshots	No scalable to other objects	No
Korsgaard *et al.* (2017) [31]		Oculus Rift CV1 + Touch OVRvision PRO	Local reality with food	Head orientation: angle in 25-30 deg	No optimal user immersion	Yes (6 subjects)
Urooj *et al.* (2018) [63]		–	Hands in the Wild	DeepLearnig (RefineNet +CRF)	Cross-dataset evaluation	No
Proposed Approach		**Samsung Gear HMD Samsung S8**	**Arms**	**Deep Learning (DeepLabv3+)**	**Re-training for Scalability**	**No**

segment egocentric arms. Then, Section V explains the experimental protocol and test datasets considered to conduct the experiments, while Section VI reports the segmentation results and the comparison with former segmentation approaches used for AV. Finally, Section VII concludes the paper with some discussions and future research lines.

II. RELATED WORKS

This section discusses the main related works on segmentation methods of local reality objects in AV. Our aim here is to conceptually categorize most relevant methods into color, depth as well as other approaches and provide a brief description thereof. In addition, Table 1 lists the main properties of these works concerning the segmentation method used, the augmented elements and the VR devices used (for information regarding state-of-the-art Sem-Seg approaches, we refer to [23]).

A. COLOR-BASED APPROACHES

One of the preliminary approaches for segmenting objects from local reality was the chroma-key, similar to the concept applied within weather forecast in television for decades. The idea is simple: given an input video with this chroma-key color presented, only pixels not sharing this color are retained. Metzger [45], one of the pioneers of the idea of AV, put forward the use of blue chroma-key, to select the user's hands from the local reality. Further, the authors pointed out the importance of having the space uniformly lit to obtain accurate results. Similarly, McGill *et al.* [44] used a green chroma-key to filter objects from the local reality (see Fig. 2 A). The particular task involved typing with a keyboard in a VR environment. For this purpose, they designed a scenario with a green chroma-key surface where the keyboard was placed. The segmentation was performed in two stages: first, both hands and the keyboard were segmented by discarding all pixels that shared the green color; then hand detection and keyboard actions were carried out using blob detection and hand markers, respectively. Although the results obtained with this simple method were almost perfect in terms of segmentation, the application itself is very limited if the local reality appearance is constrained to exhibit a certain chroma-key color. The same strategy was carried out by Zhu *et al.* [73], using a green-chroma to introduce user's body into the IE. Additionally, thanks to their real-time performance, they enabled the realistic virtual reality experience to be shared among a large number of people at the same time.

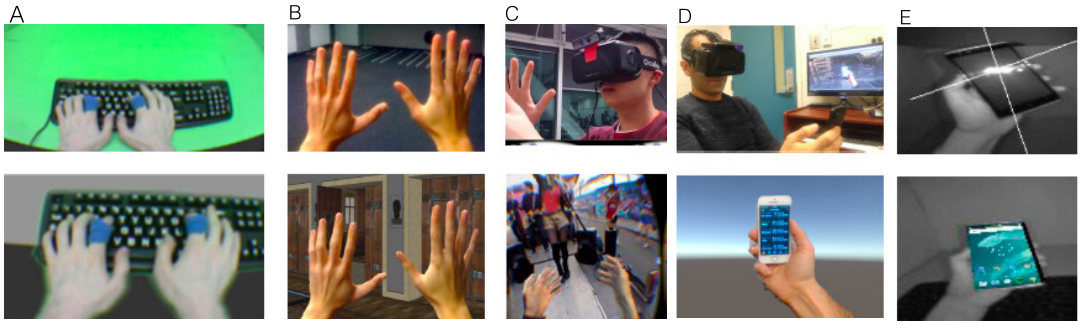

FIGURE 2. Examples of the different segmentation approaches proposed in the literature to segment local reality objects. From left to right: *A*) green chroma-key [44], *B*) skin detection [9], *C*) – *D*) depth information [1], [32] and *E*) edge detection and statistical classifier [18].

Focusing particularly on the hand segmentation problem, researchers have also proposed the use of skin detection algorithms to segment hands from local reality [9] (see Fig. 2 B). The idea behind this algorithm is the following: the local reality image is first transformed to the HSV color space, and then it is filtered out so that only pixels values that are around a certain Hue range ($\mu \pm \sigma$) are segmented. Although this approach enhanced the green chroma-key approach in the sense that local reality is not constrained anymore, some false positives may appear having similar skin color such as faces in the scene, furniture, boxes, etc. In the same work, the egocentric lower body part was also segmented with a naive floor subtraction approach. Taking the assumption that the floor appearance was uniform, the body was retained by simply filtering out all pixels not belonging to the floor. Then, a follow-up work using the same HSV-based segmentation algorithm, developed a user interface where users could adjust segmentation parameters, transparency level or even modify hand color [29].

In the same line, Perez *et al.* [49] used a YCbCr skin detection algorithm based on red chrominance, adding a transparency alpha layer to the local reality. By using less strict thresholds than those normally used for skin detection, objects with yellow and red tones with high saturation such as food were also segmented. This segmentation method allowed them to build a proof of concept of an Immersive Gastronomic Experience using Distributed Reality [64], a new type of Mixed Reality that involves capturing different realities (at least one remote in the form of 360° video and a local reality) to foster remote human communications and shared experiences.

Despite the popularity of color-based approaches and their real-time performance, they have some limitations: they require controlled physical setups, where no background objects have any of the colors included in the foreground (this is especially restrictive in traditional green chroma). Also, they are very sensitive to illumination and fail at dealing with different skin colors or with long-sleeve clothes [22].

B. DEPTH-BASED APPROACHES

Based on the idea of filtering out everything that is below a certain depth threshold value, Nahon *et al.* [48] blended into the IE not only the user's own body but also objects from the local reality and even other people. This way, self-presence is increased and also interaction and communication with other objects or people is feasible. Likewise, Lee *et al.* [32] used depth information to include the user's own body into an immersed cinema experience (see Fig.2 C). Alaee *et al.* [1] also incorporated objects which were in the distance range of $10 - 40$ cm, with the aim of interacting with the smartphone in the IE (see Fig. 2 D). More recently, Rauter *et al.* [51] implemented the same idea while estimating depth from the stereo camera of HTC VIVE Pro. They also performed some post-processing of the estimated foreground mask to address pixels with missing depth values.

Depth-based solutions are relatively simple to implement, due to the affordability of RGB-D sensors. However, such sensors have some limitations: on the one hand, depth estimation is noisy and prone to artifacts when handling near objects, specular materials, non-reachable areas, or in the shade [47]; on the other hand, RGB-D sensors have a narrow field of view which also impairs the sense of presence [32].

C. OTHER APPROACHES

Aside from the mainstream segmentation approaches, other alternatives have been proposed. Fiore et Interrante [22] built probabilistic models based on histograms for both foreground and backgrounds to segment hands, producing good results in the absence of wooden objects. Later, Desai *et al.* [18] proposed a method to segment smartphones or tablets based on two stages: 1) edge-based object detection to select the smartphone; and 2) a statistical classifier based on attributed features to decide whether the segmented object was a smartphone (see Fig.2 E). The overall goal was to interact with these devices while being immersed. This algorithm, however, was not scalable to segment other objects.

Korsgaard *et al.* [31] conducted an AV experience in which the user had to interact with real food placed in front of him. Merge between both local and remote worlds was controlled through head orientation. Every time the head was orientated in a downward angle (where food is normally placed), the local reality was visible whereas if the user looked straight ahead, the IE became visible. The main limitation of this approach is that no optimal full immersion is achieved but

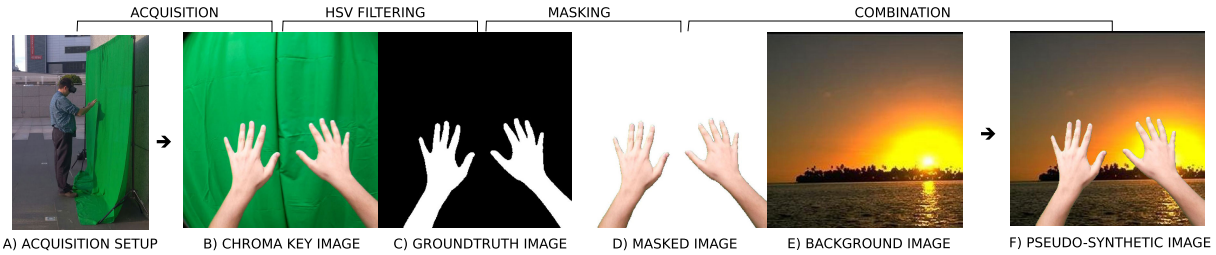

FIGURE 3. Procedure to obtain groundtruth and semi-synthetic images: through an Android app installed on the smartphone, images are recorded from the HMD perspective using a green chroma-key approach. Subsequently, we applied HSV filtering to obtain the groundtruth images. With the groundtruth image, we select the relevant information from the chroma-key image that will be later combined with a background image to form the final semi-synthetic image.

just an angle-based transition approach between the IE and the local reality.

Beyond using skin color, there are other attempts, not specifically designed for AV, to segment hands from an image-based point of view. Serra *et al.* [59] proposed a hand-crafted method for segmenting skin based on random forest superpixel classification, considering light, time, and space consistency. Although it may be seen as an evolution of color-based methods, this approach would still fail to segment arms containing clothes. There are also some attempts to detect [5] or segment hands [63] using deep learning that show the feasibility of adapting existing pre-trained models such as RefineNet or CaffeNet (slight modified versions of Alexnet). Again, these approaches are focused on segmenting hands, but not arms.

III. EgoArm DATASET

At the present time of writing, there are no databases in the literature suitable for egocentric arm segmentation. There exist some databases that are related to egocentric hand detection or segmentation but not related to the whole arm. Therefore, we introduce the *Egocentric Arm Segmentation Dataset (EgoArm)*, which is designed with a wide range of variations to maximize generalization capabilities. Table 2 describes the main features of EgoArm, containing more than 10, 000 images. We highlight that EgoArm includes images of people with different skin color, gender and it is publicly available for research purposes.[3]

TABLE 2. Features considered in EgoArm dataset.

Feature	Values
People	male (9) and female (4)
Arm pose	close hands, open palm, open dorsum, left arm, right arm
Scenario	outdoors, indoors
Outfit	outfit1, outfit2
Sleeve	long-sleeve, short-sleeve
Ethnicity	caucasian, black, mixed

Unlike other supervised learning approaches such as classification or regression, in which the required labels or groundtruth are just text labels or a few numbers defining bounding boxes, Sem-Seg labels are images where every pixel contains a particular number accounting for the class information. The acquisition of such databases is time-consuming, which represents a major problem that has already been observed by Bandini and Zariffa [6]. To overcome this issue, we propose a semi-automatic way of labeling images (see Fig. 3), composed of the following steps:

A. ACQUISITION

First, the user wears a Gear VR Samsung headset with a Samsung-S8 smartphone attached to the device and situated himself/herself in front of a chroma-key backdrop, as can be seen in Fig. 3 A. An Android application is used to record videos of 30 fps from the smartphone frontal camera. Unlike other segmentation datasets, we recorded videos at a resolution of 720 × 720 pixels to target the high-resolution requirements of VR applications (Fig. 3 B). Each session is designed to record videos with a particular configuration in terms of people, scenario, outfit, and sleeve. A recorded assistant ensures that, at each session, videos from the five different arm poses are recorded.

B. HSV FILTERING

With the recorded chroma-key videos (see Fig. 3 B), an HSV-based filter is applied to obtain the foreground images (see Fig. 3 C), as follows:

$$f(x,y)=\begin{cases}1 & \text{if } H(x,y)\leq h_1 \wedge H(x,y)\geq h_2 \wedge S(x,y)\geq s_1\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

being h_1, h_2 and s_1 set to 0.22, 0.45, and 0.20, respectively (values obtained by empirical testing) and $H(x,y)$ and $S(x,y)$ being the Hue and Saturation channel images, respectively. To prevent high similarity, images are selected every 5 frames. Additionally, some morphological operations are applied to delete noisy areas (see Fig. 3 C).

C. MASKING

Before creating the semi-synthetic image, the chroma-key image is masked with the groundtruth image to get the area of interest (arms in our proposed approach, see Fig.3 D).

[3] https://cloud.proinnovation.es/index.php/s/tekqtneGXgrUgFD

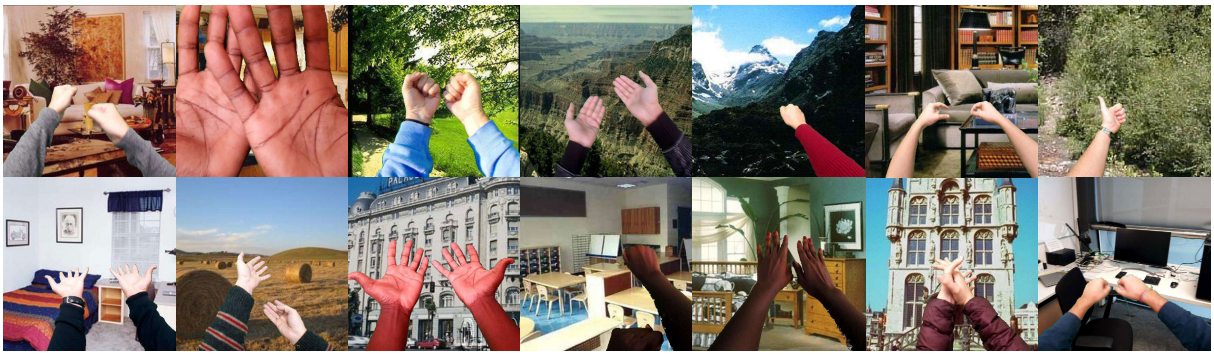

FIGURE 4. Example images of the EgoArm dataset showing a wide variety in terms of different subjects, gender, arms position, scale, clothes, skin color, illumination and background.

D. COMBINATION

Semi-synthetic images (Fig.3 F) are created combining background (Fig.3 E) with chroma-key images (Fig. 3 B) masked with foreground images (Fig.3 D). In this work, natural background images are obtained from the MIT Scene Parsing Benchmark [72]. Among the whole set of 20, 210 images, we select those which hold $height = width$ and then reshape it to 720×720, resulting in a subset of 3, 697 different background images. These backgrounds contain indoor scenes related to houses, public spaces, commercial places as well as outdoor scenes such as landscapes, beaches, mountains, etc. As a final post-processing, we discarded those pairs of *groundtruth* and semi-synthetic images with some false positives in the groundtruth. Fig.4 shows examples of the variability of these images.

IV. EGOCENTRIC ARM SEGMENTATION

Accurate and robust arm segmentation is vital to achieve enhanced self-perception in MR. DL-based approaches have outperformed conventional approaches in diverse computer vision tasks whenever available training data reflects real-world scenarios. This clearly motivates the development of a novel arm segmentation system based on deep learning in order to overcome the disadvantages existed in traditional approaches (see Section II). Convolutional Neural Networks (CNN) achieved the state-of-the-art results for supervised classification and detection tasks [28]. CNN are composed of different types of hidden layers: Convolutional, Rectifier Linear Unit, Pooling and Fully Connected (FC). FC are the final layers of CNNs that, along with the classification layer, hold the output (having the same size as the number of objects to classify). In 2015, Long *et al.* proposed Fully Convolutional Networks (FCN): a modification of CNN architectures that reached state-of-the-art performance in Sem-Seg problems. Concretely, they replaced FC layers by fully convolutional ones to preserve the spatial dimension while keeping the class identity information [42]. Another important key component aside from the encoding subnetwork here is the decoding subnetwork, which is placed after the fully convolutional layers and is in charge of up sampling the class spatial map up to the original input size.

A. CONSIDERED SEM-SEG NETWORKS

Our hypothesis, confirmed also by previous work [56], is that segmentation networks trained for TPV fail when segmenting from egocentric vision. Indeed, egocentric vision has the advantage that objects tend to appear at the center of the image, but also the challenge of the camera moving with the human body, which creates fast movements and sudden illumination changes.

Due to the relatively small size of EgoArm (in comparison with datasets aimed to train architectures from scratch such as ImageNet, Pascal VOC, etc.), we took the decision to apply transfer learning from existing Sem-Seg architectures. The first Sem-Seg architecture considered was the FCN, proposed by Long *et al.* and originally trained for the PASCAL VOC 2011 segmentation challenge. We conducted extensive experiments to find the best training parameters for fine-tuning the FCN architecture with the EgoArm database. Still we observed that results were not accurate enough for the 720×720 required resolution.

The next Sem-Seg deep architecture we considered was DeepLab, originally proposed in 2017. In particular, among their different updated versions [11]–[13], we chose DeepLabv3+ [13] due to: *i*) the use of the ResNet pre-trained model, replacing the former VGG-16 pre-trained model; *ii*) the use of *a-trous* convolutions, that allow dense feature extraction taking context into account without increasing the number of parameters; *iii*) the use of *a-trous* spatial pyramid pooling module to robustly segment objects at multiple scales; and *iv*) the use of a decoder module and short-cut connections [57] to refine the segmentation results [4], especially along object boundaries. The fact that this architecture was very deep at the encoding subnetwork and deeper than the existing approaches in the decoding subnetwork gave us the idea that it could segment accurately high-resolution egocentric images. For a better understanding, we decided to use the following two different Sem-Seg networks:

- **DeepLabv3+**: our idea here is to use the original DeepLabv3+ to segment egocentric arms and confirm our hypothesis. This original network was trained using the PASCAL VOC database, so arms were segmented as people class.

- **DeepLabv3+ using EgoArm**: we applied transfer learning using images from our novel EgoArm dataset so that the network segments two classes: arms and background. In order to have a more gender-balanced dataset, we discarded 4 male subjects, having a total of 11, 561 images.

V. EXPERIMENTAL PROTOCOL

The main motivation for not using a subset of EgoArm as the validation set, was to check how well a network trained with semi-synthetic images generalizes with real egocentric images. Among the public real egocentric datasets, GTEA Gaze+ was the largest one and more similar to the arm segmentation task [5]. It contains 1, 115 images of egocentric arms performing actions in a kitchen, with a very cluttered environment. In this dataset, groundtruth is related to the skin but no clothes were presented in the images.

Training was done using two GPU GTX-1080 Ti with $12GB$ RAM each. Batch size was set to 4 due to the large size of the training images (720×720). An exhaustive set of experiments following grid search strategies were conducted, monitoring validation performance over the GTEA Gaze+. The final training of the DeepLabv3+ EgoArm was achieved using stochastic gradient descent, an initial learning rate of $1e-3$, a final learning rate of $1e-6$, 2 epochs, 7, 500 maximum number of iterations for reducing the learning rate, and weight decay of $1e-5$.

A. TESTS

In order to assess the generalization capabilities of our algorithm, we performed the evaluation on the following different public datasets. Fig. 5 describes the associated heatmaps, to give an idea of the type of groundtruth and the average position of hands/arms[4]:

- **EDSH (groundtruth related to skin)** [36]: EDSH2 and EDSH kitchen are the test videos of EDSH, and contain indoor and outdoor scenes with large variations of illumination, mild camera motion induced by walking and climbing stairs with just one user. They provide 104 and 197 segmentation masks for EDSH2 and EDSHK, respectively.
- **EgoHands (groundtruth related to hands)** [5]: it contains 48 Google Glass videos of interactions between two people playing board games (one with FPV, and the other with TPV). In order to reduce redundancy and computational load, we created a subset of this dataset, by selecting 10 images per each of the 48 different videos, resulting in a total of 480 images.
- **Ego Youtube Hands (groundtruth related to hands)** [41]: it contains 3 egocentric videos from daily activities.

[4]There were also other datasets available in the literature that we discarded for different reasons. For instance, the EPIC-KITCHENS dataset does not provide segmentation masks [16]; the Egocentric Gesture Recognition dataset [10] only provides segmentation masks for chroma-key hand gesture images; and Keyboard Hand Dataset (KBH, [66]) was not found available for research purposes.

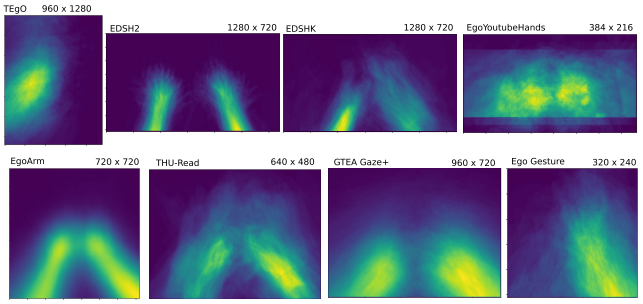

FIGURE 5. Heatmaps depicting hand/arm position in the validation and test datasets considered. From up to bottom and left to right: TeGO, EDSH2, EDSHK, Ego Youtube Hands, EgoArm, THU-Read, GTEA Gaze+, and Ego Gesture.

Among the entire set of 1, 032 images, we created a subset including images showing hands and arms, resulting in a total of 689 frames.

- **TEgO database (groundtruth related to skin)** [33]: to test the robustness against black skin color, we report results using the test set pertaining to subject $B1$ (which has black skin), composed of different subsets of images under different illumination (normal and extreme) and background conditions (vanilla or in the wild).
- **FPAB dataset (no groundtruth available)** [24]: provides both color and depth images from egocentric images. As their original purpose was to infer hand pose, people wear a mo-cap device on the right hand. As the color and depth images were extracted from different sensors and at different positions, it was very difficult to create a common groundtruth, so we do not report empirical results, but only visual examples.
- **Ego Gesture (groundtruth related to arms)** [70]: contains egocentric color and depth videos acquired from RealSense SR300. It includes 83 different hand gestures from 50 different subjects and 6 different scenarios (e.g. indoors, outdoors, illumination, static clutter background, dynamic background, walking, etc.). As the groundtruth was related to hand gestures, we manually labeled the arm segmentation masks of a representative subset of 277 images (by sampling approximately one image per subject and scenario).
- **THU-Read (groundtruth related to arms)** [61]: is created for egocentric action recognition from RGB-D data. Concretely, there are up to 40 different actions, each related to particular objects, performed by 8 different users. Due to the lack of pixel-wise labeling for both RGB and depth images, we also manually labeled the segmentation masks of a representative subset of 203 images, ensuring variability of actions and users.

B. PERFORMANCE METRIC

Empirical results are given in terms of Jaccard Index, also known as Intersection over Union (IoU), defined as:

$$mIoU = \frac{1}{k}\sum_{i=1}^{k} IoU_i = \frac{1}{k}\sum_{i=1}^{k}\left[\frac{TP}{TP+FP+FN}\right]_i \quad (2)$$

TABLE 3. Segmentation results in terms of Intersection over Union (IoU) for different egocentric segmentation datasets. The considered segmentation algorithms for AV are: 1) color, 2) baseline Deeplabv3+ using the person class; 3) Deeplabv3+ using the proposed EgoArm dataset. GTEA Gaze+ is our validation dataset. The reader should keep in mind that there is discrepancy between the available groundtruth (related to hands or skin) with the arm concept. Due to that, *MissRate* is also reported. Bold indicates best *IoU* for a given dataset. Results format follows *IoU* (*MissRate*).

Database	# Images	Color	DeepLabv3+	DeepLabv3+ EgoArm
GTEA Gaze+ [5]	1115	15.47 (4.13)	40.57 (50.68)	**60.75** (7.50)
EDSH2 [36]	104	52.62 (16.93)	67.56 (13.13)	**74.04** (8.30)
EDSHK [36]	197	42.24 (11.23)	52.50 (22.91)	**56.61** (8.10)
EgoHands [5]	480	29.09 (16.19)	26.65 (18.69)	**33.52** (17.85)
Ego Youtube Hands [41]	689	**22.11** (34.35)	20.64 (49.67)	20.80 (35.71)
Ego Gesture [70]	277	43.96 (42.15)	43.56 (43.28)	**67.94** (18.54)
THU-Read [61]	203	25.62 (42.23)	51.35 (22.00)	**57.75** (6.33)
TEgO Vanilla [33]	190	2.45 (77.43)	12.03 (84.36)	**46.84** (13.63)
TEgO Vanilla illu		2.96 (72.94)	14.81 (81.02)	**48.30** (9.20)
TEgO Wild		16.45 (47.26)	16.50 (74.22)	**37.25** (10.23)
TEgO Wild ilu		18.09 (57.30)	16.34 (79.55)	**54.76** (8.28)
Overall	–	IoU 24.64 ± 15.51 $MissRate$ 38.37 ± 23.47	IoU 32.95 ± 18.07 $MissRate$ 49.04 ± 26.10	IoU 50.97 ± 14.87 $MissRate$ 13.06 ± 8.17

where k is the number of classes (in our case $k = 2$: arms and background). IoU is first computed independently for each class and then averaged to have mean Intersection over Union $mIoU$. IoU measures the ratio between intersection between groundtruth and predicted segmentation masks over their union. True Positives (TP) are the number of pixels belonging to the target class which were successfully predicted. False Positives (FP) account for the number of pixels which were wrongly predicted to belong to the target class. False Negatives (FN) counts the number of pixels belonging to the target class but were not correctly predicted. Due to the great imbalance of pixels belonging to arm and background per image, we report exclusively IoU pertaining to the arm class, in the range 0-100%.

As grountruth of the available test datasets presented in Section V-A is only related to hands or skin, but not clothes, reported IoU is underestimated. This means that clothes, even when segmented by our proposed method, count as FP and thus, reduce the IoU. For further understanding, we also reported $MissRate = \frac{FN}{FN+TP}$, also in the range 0-100%. $MissRate$ accounts for the percentage of pixels belonging to the hand/skin class wrongly classified as background. Therefore, the smaller, the better.

VI. RESULTS

In the following we provide an extensive quantitative and qualitative assessment of our proposed deep segmentation network, in comparison with former segmentation approaches used in AV, namely color and depth. Concretely, Section VI-A starts the discussion introducing the baseline performance of color-based segmentation in comparison with the baseline DeepLabv3+ deep network; Section VI-B follows it comparing the performance of baseline DeepLabv3+ with the one trained with EgoArm. Last, Section VI-C will compare deep-based with respect to depth approaches. Table 3 and Table 4 present *IoU* and *MissRate* values for RGB and RGB-D datasets, respectively.

A. COLOR PERFORMANCE

Color-based segmentation is applied using an HSV filtering similar to Equation 1. As can be seen from Table 3, color-based segmentation achieves similar or worse results than the baseline DeepLabv3+. Concretely, there is an absolute improvement from 10.00% to 25.00% IoU when replacing color-based to DeepLabv3+ for GTEA Gaze+, EDSH, THU-Read and TEgO datasets. From a high-level perspective, this is expected since deep learning algorithms, unlike color-based ones, consider additional information beyond color such as shapes, texture, and other complex information. Due to this reason, performance is also hindered when there are objects in the scene which share the skin color; notice the very bad performance of GTEA Gaze+ or THU-Read databases due to their yellowish/reddish scene appearance (see Fig.6 A and J). Also, results reported from the TEgO database show that relying exclusively on color is not an appropriate method when users from different ethnicities are involved (see Fig.7). Besides, this method fails at segmenting clothes (see Fig.6 G and L and Fig.8 A).

In the case of Ego Gesture, there are no observable performance differences between the color and the baseline DeepLabv3+, according to results reported in Table 3. However, when assessing those results per scene

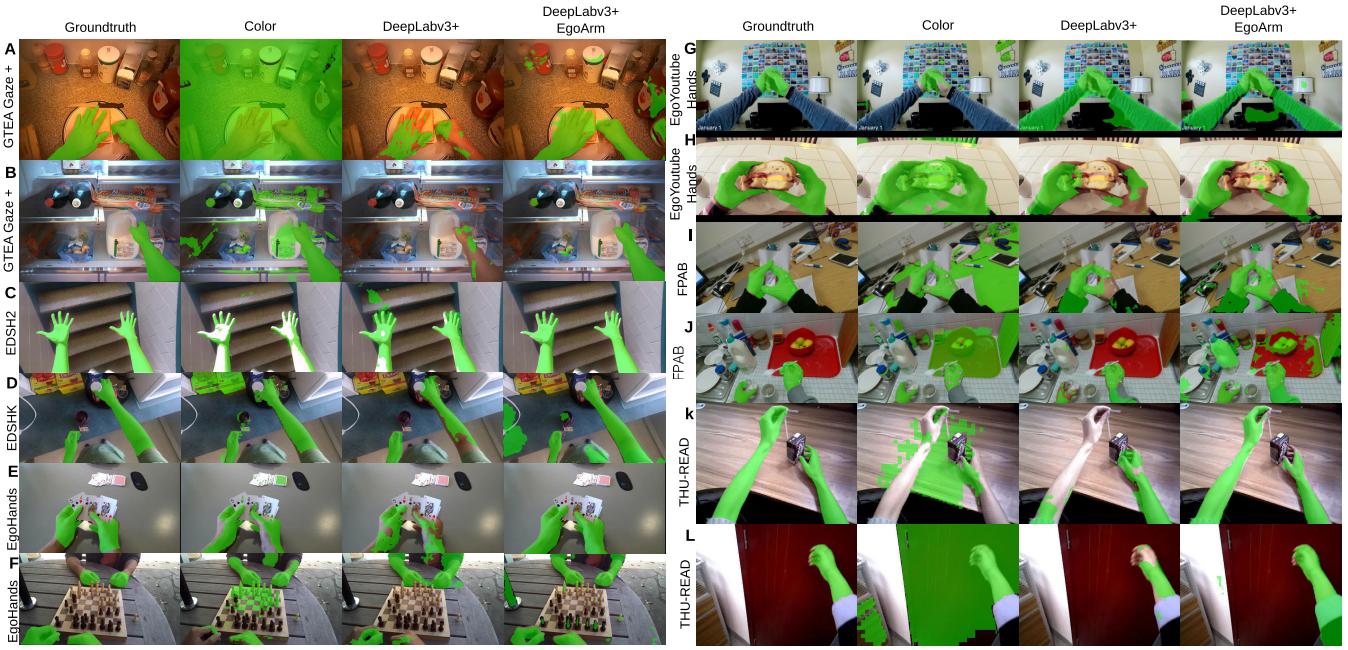

FIGURE 6. Segmentation examples of the different segmentation methods explored for GTEA GAZE+, EDSH, EgoHands, Ego Youtube Hands,TeGO, FPAB and THU-Read. The first column refers to the groundtruth defined in each case. Notice that FPAB do not contain groundtruth images, we have manually labeled these two examples.

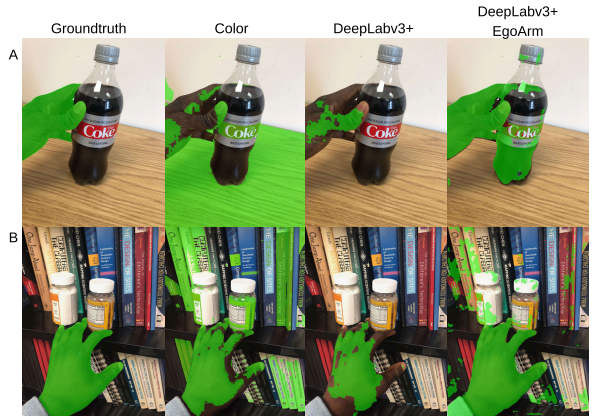

FIGURE 7. Samples from TeGO. Notice that groundtruth is related to the skin.

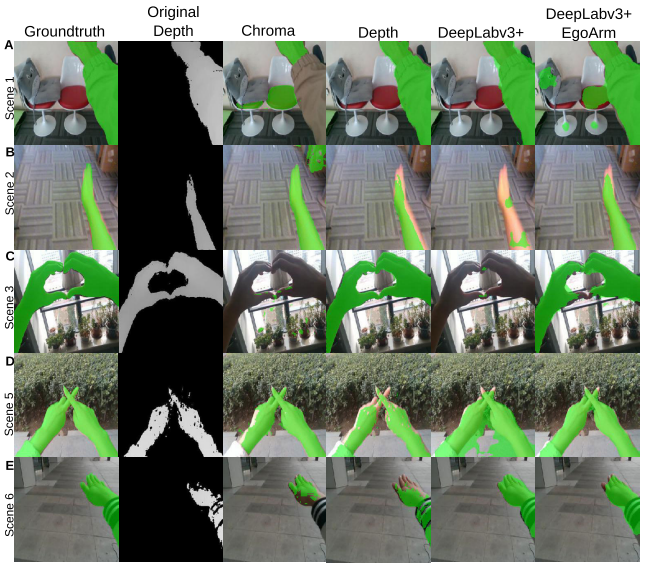

FIGURE 8. Comparison results using Ego Gesture dataset, composed of RGB and depth egocentric images in 6 different scenes. Notice that groundtruth is related to arms.

(see Table 4), we observe that: *i*) both color and DeepLabv3+ are severely affected by extreme illumination conditions (Scene3, see Fig.8 C); *ii*) color is more robust than DeepLabv3+ in both dynamic or walking indoor scenarios where movement can produce some blur effect (around 10.00% average absolute improvement when using color rather than DeepLabv3+ for Scene2 and Scene4, see Fig.8 B), and *iii*) that DeepLabv3+ outperforms color when good illumination is available (Scene5 and Scene6, see Fig.8 D-E) or there is a controlled background.

B. DEEP PERFORMANCE

Concerning the behavior of the two deep Sem-Seg networks, we observe the general superiority of the network trained with EgoArm in comparison with the baseline DeepLabv3+. This observation validates our hypothesis of the convenience of a database more similar to the real application.

We observe a slight, moderate or large improvement, depending on the particular dataset. Slight improvement is observed for EgoHands (6.87% absolute improvement) and no improvement is observed for the Ego Youtube Hand datasets. In both cases, results are very poor due to the

TABLE 4. Comparison results using Ego Gesture dataset, composed of RGB and depth egocentric images in 6 different scenes, and THU-READ dataset. Results format follows *IoU* (*MissRate*). Bold indicates best *IoU*.

Scene	Color	Depth	DeepLabv3+	DeepLabv3+ EgoArm
Ego Gesture Scene1 Indoor Clutter Background	40.37 (31.86)	**77.57** (14.42)	44.64 (39.72)	49.26 (23.73)
Ego Gesture Scene2 Indoor Dynamic Background	48.72 (28.95)	**75.00** (20.01)	38.39 (50.78)	70.33 (21.37)
Ego Gesture Scene3 Indoor Toward Windows	21.40 (70.54)	**73.91** (20.64)	25.48 (71.75)	67.32 (21.54)
Ego Gesture Scene4 Indoor Walking	47.25 (37.13)	**76.54** (19.25)	35.98 (37.47)	62.93 (20.05)
Ego Gesture Scene5 Outdoor Dynamic Background	49.54 (46.79)	72.57 (24.22)	57.50 (28.53)	**77.09** (10.76)
Ego Gesture Scene6 Outdoor Walking Dynamic Background	56.69 (38.78)	69.79 (28.21)	61.58 (29.83)	**79.97** (12.22)
THU-READ subset	25.62 (42.23)	19.63 (52.54)	51.35 (22.00)	**57.75** (6.33)

groundtruth being related just to hands despite most images show whole arms with or without clothes (see Fig.6 G and F). Also, in the case of EgoHands (see Fig.6 F), the majority of images present both FPV and TPV arms. Moreover, TPV arms occupy a larger surface than FPV ones. As the network trained with EgoArm focuses on FPV arms, it makes sense that the improvement on such database is limited. In what concerns Ego Youtube Hands images, we assume that their low resolution (384 × 216) along with their uncontrolled and cluttered environment makes the segmentation specially challenging.

A slight gain is observed with EDSH2 when including EgoArm (6.50% absolute improvement). We believe this is because most test images just contain arms but not clothes, but also because background and hand position are controlled (e.g. fingers are very well separated). For the more uncontrolled EDSHK, there is a larger improvement specially in terms of *MissRate* (from 22.91% to 8.10%) between the baseline DeepLabv3+ and the one trained with EgoArm. Among EDSHK images, it is very frequent to encounter arms with clothes, which are not considered part of the groundtruth (see Fig.6 D). Therefore, part of *FP* is related to the clothes side of the arm.

Moderate enhancement is encountered for the GTEA Gaze+ and Ego Gesture in the range of 15.00% to 25.00% absolute improvement of *IoU*, but also in THU-READ, considering the decrease of *MissRate* from 22.00% to 6.33%. As these datasets are purely egocentric, it is more noticeable the gain when using the EgoArm (see Fig.6 A-B, Fig.6 K-L or Fig.8 A-B). Having a more in-depth look to the *IoU* per scene reported in Table 4, it is observed a huge improvement of DeepLabv3+ trained on EgoArm in all scenes and notably in outdoors scenarios, which apart from the nature of the scenarios, might benefit from uniform and good illumination. Lastly, a considerable increase in performance is noticed with the different subsets from TEgO dataset (in the range of 20-40% of absolute improvement). The main reason behind it resides on the diversity of skin colors presented in the EgoArm.

After having a visual inspection to the images, we notice that in some cases, the DeepLabv3+ EgoArm network generates some false positives from background items. We hypothesize this has to do with the huge variability (both in terms of color, shapes, textures) of backgrounds, which might not be fully retained by the network parameters.

C. COMPARISON WITH DEPTH

Here we provide a detailed analysis of the results obtained with depth in comparison with color-based or deep-based segmentation, using the Ego Gesture and THU-READ RGB-D subsets described in Section V-A.

As stated in Section II, segmentation based on depth implies the selection of all objects that are below a particular distance threshold d_1 from the sensor, as defined by the following equation:

$$f(x,y) = \begin{cases} 1 & \text{if } 1 \leq Depth(x,y) \leq d_1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where x and y refer to the pixel position in the image. In our experimental framework, we set d_1 to a value greater than 0 to discard unknown depth pixels.

It is clearly visible from Table 4 that the segmentation based on depth is more uniform across the different indoor scenes of Ego Gesture than deep- or color-based approaches. A slight drop in the depth performance is shown in outdoor scenarios (see Fig.8 D-E), possibly because signal light[5] is much weaker than ambient sunlight. As a result, depth estimation works fine in Ego Gesture as the considered scenes

[5]Texture being projected in infrared to compute depth through disparity.

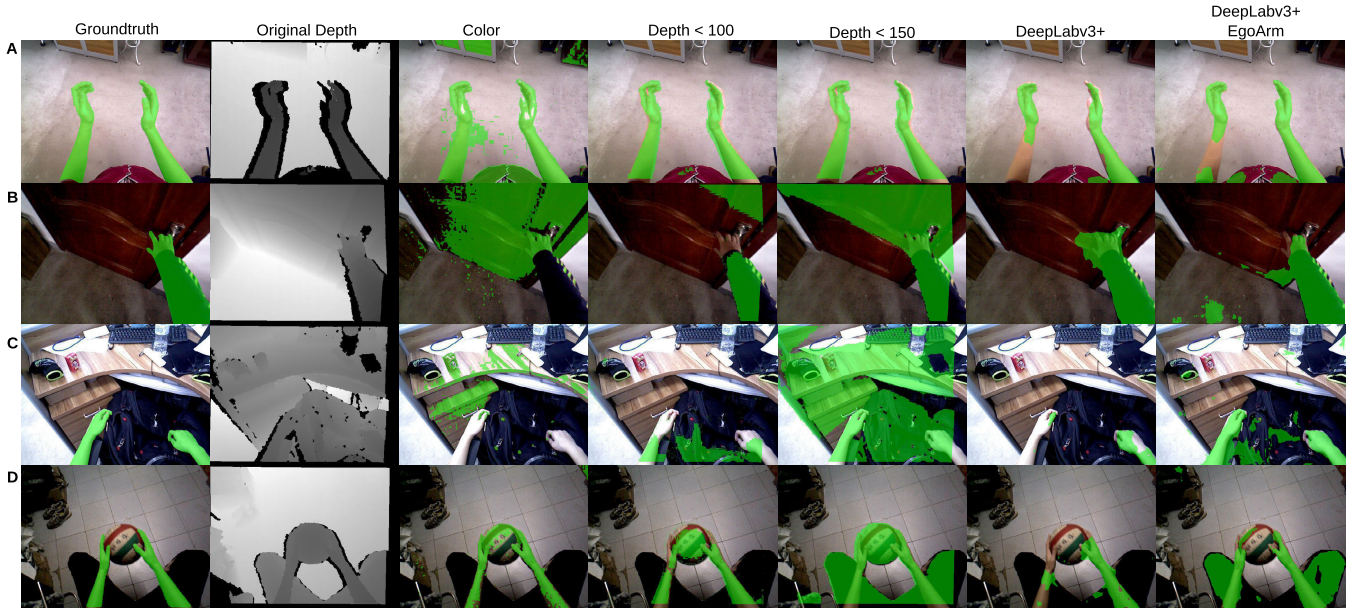

FIGURE 9. Comparison results using THU-READ dataset, composed of 203 RGB and depth egocentric images in cluttered and challenging scenarios. Notice that groundtruth is related to arms.

avoid all the critical scenarios for RGB-D sensors [47]: *i*) hands/arms are always within the distance range of the camera (and never closer) and no other object is in such range; *ii*) hands are fully visible from both infrared sensors, and *iii*) they never cast shadows from the infrared emitter.

In contrast, images of the RGB-D THU-READ subset showcase more challenging scenarios: arms are always surrounded by other objects or furniture. Thus, resulting depth images are much noisier due to occlusions when estimating depth through disparity. It also presents a narrower field of view with respect to the RGB sensor (see Fig.9). As reported by Table 4, depth performs very poorly in these cluttered scenarios, reaching an IoU of 19.63%, in contrast to 57.75% achieved by DeepLabv3+ EgoArm. Indeed, a threshold-based segmentation have many limitations: arms can be partially segmented (see Fig. 9 B) additional objects segmented beyond the desired ones (see Fig.9 C-D). In some situations, having a dynamic threshold could solve some of the limitations of depth, but still it would not be trivial, especially in those cases in which arms and other objects share the same distance range. Also, depth sensors have a minimum distance at which depth can be estimated. This also hinders the segmentation in a situation where arms or objects are very close to the sensor.

Based on the aforementioned results, depth works well at segmenting arms in isolated areas (see Fig.9 A) without any other object at the same distance range, but its performance severely degrades in outdoor and cluttered scenarios. Moreover, there are recent studies exploring deep learning to enhance depth maps (also known as depth completion), that suggest that there is still a large room for improvement in this area [40], [43], [71].

D. COMPUTATION TIME

Given a 720×720 image, segmenting it with color, depth, and deep approaches would take $2.9ms$, $700\mu s$[6] and $74ms$, respectively using a PC Intel Xeon ES-2620 V4 @ 2.1Ghz with 32 GB powered with 2 GPU GTX-1080 Ti with $12GB$ RAM. Our deep implementation achieves about 15 fps, which is 4 to 6 times slower than what it would be desirable for a smooth AV system. However, it falls within the right order of magnitude, so it is just a matter of algorithm optimization and hardware improvement that the Sem-Seg approach can work in real time. In practice, it would imply either having the HMD attached to a resourceful computer or offloading computation to the edge cloud [19], [37].

VII. CONCLUSION

Egocentric perception has attracted the interest of the AV community due to egocentric cameras being incorporated to VR headsets. In particular, egocentric image segmentation can provide new features such as increasing sense of presence, if the user's body parts are segmented, or interactivity with the local reality, if nearby objects (e.g. laptop, mobiles) are segmented. In this work, we aimed to shed some light on segmentation methods for AV beyond the traditional approaches based on color or depth. Our main contributions can be summarized as follow: *i*) We first conduct a **comprehensive review** of previous studies proposing segmentation methods for AV, focusing on the method, the objects segmented, and the VR devices used; *ii*) we create the **EgoArm dataset** composed of more than 10, 000

[6]This does not include the time required to generate the depth map from the stereoscopic images.

semi-synthetic images, **demographically inclusive** (variations of gender, skin color), and **open for research purposes**; *iii*) we provide a very detailed description on how to generate semi-synthetic images and the automatic method to generate pixel wise labeling, which will **help future researchers to create their own custom datasets** at a low cost; *iv*) the use of deep learning for the first time for segmenting arms in AV; *v*) we report results and provide insights while testing our proposed network trained with the semi-synthetic EgoArm dataset with **8 real egocentric datasets**: GTEA Gaze+, EDSH, Ego Youtube Hands, EgoHands, FPAB, THU-Read, TEgO, and Ego Gesture, providing comparisons with color- and depth-based segmentations. Results have proven the effectiveness of EgoArm for arm segmentation, boosting the average *IoU* from 25.00% reached with color or from 31.35% reached with the baseline DeepLabv3+ network, up to 50.00% *IoU*. Besides, these segmentation networks are more robust than color-based approaches when dealing with illumination changes, segmenting clothes or arms with different skin color, etc. In comparison with depth, deep-based segmentation algorithms are also more robust in outdoor or cluttered scenarios. Segmentation based exclusively on depth works reasonably well for isolated scenarios. This encourages us to foresee a potential use of depth information to complement the training deep networks based on RGB; it might help to mitigate the problem of false positives already mentioned. Potential future research lines could be:

- Make it work on real time by exploring the tradeoff between inference time and accuracy of shallower architectures such as [38], [68].
- Explore multimodal approaches based on RGB and depth input images.
- Evaluate the presence and sense of embodiment properties of the proposed segmentation method using standardized questionnaires [26], [60], [67].
- Extend the proposed segmentation method to other classes, such as human body or some specific objects, and allow interaction with virtual objects by tracking human body parts [30].

REFERENCES

[1] G. Alaee, A. P. Deasi, L. Pena-Castillo, E. Brown, and O. Meruvia-Pastor, "A user study on augmented virtuality using depth sensing cameras for near-range awareness in immersive VR," in *Proc. IEEE VR's 4th Workshop Everyday Virtual Reality (WEVR)*, Mar. 2018, pp. 1–10.

[2] F. Argelaguet, L. Hoyet, M. Trico, and A. Lecuyer, "The role of interaction in virtual embodiment: Effects of the virtual hand representation," in *Proc. IEEE Virtual Reality (VR)*, Mar. 2016, pp. 3–10.

[3] L. Aymerich-Franch, R. F. Kizilcec, and J. N. Bailenson, "The relationship between virtual self similarity and social anxiety," *Frontiers Human Neurosci.*, vol. 8, p. 944, Nov. 2014.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[5] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1949–1957.

[6] A. Bandini and J. Zariffa, "Analysis of the hands in egocentric vision: A survey," 2019, *arXiv:1912.10867*. [Online]. Available: http://arxiv.org/abs/1912.10867

[7] A. Betancourt, P. Morerio, C. S. Regazzoni, and M. Rauterberg, "The evolution of first person vision methods: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 744–760, May 2015.

[8] M. Bolaños, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 77–90, Feb. 2017.

[9] G. Bruder, F. Steinicke, K. Rothaus, and K. Hinrichs, "Enhancing presence in head-mounted display environments by visual body feedback using head-mounted cameras," in *Proc. Int. Conf. CyberWorlds*, Sep. 2009, pp. 43–50.

[10] T. Chalasani, J. Ondrej, and A. Smolic, "Egocentric gesture recognition for head-mounted AR devices," in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2018, pp. 109–114.

[11] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[12] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[13] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018, *arXiv:1802.02611*. [Online]. Available: http://arxiv.org/abs/1802.02611

[14] H. Cheng, L. Yang, and Z. Liu, "Survey on 3D hand gesture recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 9, pp. 1659–1673, Sep. 2016.

[15] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A survey on activity detection and classification using wearable sensors," *IEEE Sensors J.*, vol. 17, no. 2, pp. 386–403, Jan. 2017.

[16] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 720–736.

[17] A. Garcia del Molino, C. Tan, J.-H. Lim, and A.-H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 65–76, Feb. 2017.

[18] A. P. Desai, L. Pena-Castillo, and O. Meruvia-Pastor, "A window to your smartphone: Exploring interaction and communication in immersive VR with augmented virtuality," in *Proc. 14th Conf. Comput. Robot Vis. (CRV)*, May 2017, pp. 217–224.

[19] M. Erol-Kantarci and S. Sukhmani, "Caching and computing at the edge for mobile augmented reality and virtual reality (AR/VR) in 5G," in *Ad Hoc Networks*. Cham, Switzerland: Springer, 2018, pp. 169–177.

[20] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 407–414.

[21] T. Feuchtner and J. Müller, "Extending the body for interaction with reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 5145–5157.

[22] L. P. Fiore and V. Interrante, "Towards achieving robust video selfavatars under flexible environment conditions," *Int. J. Virtual Reality*, vol. 11, no. 3, pp. 33–41, 2012.

[23] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. Garcia-Rodriguez, "A review on deep learning techniques applied to semantic segmentation," 2017, *arXiv:1704.06857*. [Online]. Available: http://arxiv.org/abs/1704.06857

[24] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with RGB-D videos and 3D hand pose annotations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 409–419.

[25] M. Gonzalez-Franco, P. Abtahi, and A. Steed, "Individual differences in embodied distance estimation in virtual reality," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2019, pp. 941–943.

[26] M. Gonzalez-Franco and T. C. Peck, "Avatar embodiment. Towards a standardized questionnaire," *Frontiers Robot. AI*, vol. 5, p. 74, Jun. 2018.

[27] E. Gonzalez-Sosa, P. Perez, R. Kachach, J. J. Ruiz, and A. Villegas, "Towards self-perception in augmented virtuality: Hand segmentation with fully convolutional networks," in *Proc. 39th Annu. Eur. Assoc. Comput. Graph. Conf., Posters*, 2018, pp. 9–10.

[28] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*, vol. 1. Cambridge, MA, USA: MIT Press, 2016.

[29] T. Gunther, I. S. Franke, and R. Groh, "Aughanded virtuality—The hands in the virtual environment," in *Proc. IEEE Virtual Reality (VR)*, Mar. 2015, pp. 157–158.

[30] H. Khan, G. Lee, S. Hoermann, R. Clifford, M. Billinghurst, and R. W. Lindeman, "Evaluating the effects of hand-gesture-based interaction with virtual content in a 360 movie," in *Proc. Int. Conf. Artif. Reality Telexistence, Eurograph. Symp. Virtual Environ.*, 2017, pp. 1–8.

[31] D. Korsgaard, N. C. Nilsson, and T. Bjørner, "Immersive eating: Evaluating the use of head-mounted displays for mixed reality meal sessions," in *Proc. IEEE 3rd Workshop Everyday Virtual Reality (WEVR)*, Mar. 2017, pp. 1–4.

[32] G. A. Lee, J. Chen, M. Billinghurst, and R. Lindeman, "Enhancing immersive cinematic experience with augmented virtuality," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR-Adjunct)*, Sep. 2016, pp. 115–116.

[33] K. Lee and H. Kacorri, "Hands holding clues for object recognition in teachable machines," in *Proc. CHI Conf. Hum. Factors Comput. Syst. CHI*, 2019, pp. 1–12.

[34] K. M. Lee, "Presence, explicated," *Commun. Theory*, vol. 14, no. 1, pp. 27–50, Feb. 2004.

[35] C. Li and K. M. Kitani, "Model recommendation with virtual probes for egocentric hand detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2624–2631.

[36] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3570–3577.

[37] E. Li, Z. Zhou, and X. Chen, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in *Proc. Workshop Mobile Edge Commun. - MECOMM*, 2018, pp. 31–36.

[38] H. Li, P. Xiong, H. Fan, and J. Sun, "DFANet: Deep feature aggregation for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9522–9531.

[39] R. Li, Z. Liu, and J. Tan, "A survey on 3D hand pose estimation: Cameras, methods, and datasets," *Pattern Recognit.*, vol. 93, pp. 251–272, Sep. 2019.

[40] R. Li, S. Wang, Z. Long, and D. Gu, "UnDeepVO: Monocular visual odometry through unsupervised deep learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 7286–7291.

[41] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 287–295.

[42] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[43] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 3288–3295.

[44] M. McGill, D. Boland, R. Murray-Smith, and S. Brewster, "A dose of reality: Overcoming usability challenges in VR head-mounted displays," in *Proc. 33rd Annu. ACM Conf. Hum. Factors Comput. Syst. CHI*, 2015, pp. 2143–2152.

[45] P. J. Metzger, "Adding reality to the virtual," in *Proc. IEEE Virtual Reality Annu. Int. Symp.*, Sep. 1993, pp. 7–13.

[46] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays," *IEICE Trans. Inf. Syst.*, vol. 77, no. 12, pp. 1321–1329, Dec. 1994.

[47] G. Moyà-Alcover, A. Elgammal, A. Jaume-i-Capó, and J. Varona, "Modeling depth for nonparametric foreground segmentation using RGBD devices," *Pattern Recognit. Lett.*, vol. 96, pp. 76–85, Sep. 2017.

[48] D. Nahon, G. Subileau, and B. Capel, "'Never blind VR' enhancing the virtual reality headset experience with augmented virtuality," in *Proc. IEEE Virtual Reality (VR)*, Mar. 2015, pp. 347–348.

[49] P. Perez, E. Gonzalez-Sosa, R. Kachach, J. Ruiz, I. Benito, F. Pereira, and A. Villegas, "Immersive gastronomic experience with distributed reality," in *Proc. IEEE 5th Workshop Everyday Virtual Reality (WEVR)*, Mar. 2019, pp. 1–4.

[50] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2015.

[51] M. Rauter, C. Abseher, and M. Safar, "Augmenting virtual reality with near real world objects," in *Proc. IEEE Conf. Virtual Reality 3D User Interface (VR)*, Mar. 2019, pp. 1134–1135.

[52] H. Regenbrecht, T. Lum, P. Kohler, C. Ott, M. Wagner, W. Wilke, and E. Mueller, "Using augmented virtuality for remote collaboration," *Presence, Teleoperators Virtual Environ.*, vol. 13, no. 3, pp. 338–354, Jun. 2004.

[53] X. Ren and M. Philipose, "Egocentric recognition of handled objects: Benchmark and analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2009, pp. 1–8.

[54] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert, "The perception of egocentric distances in virtual environments–A review," *ACM Comput. Surveys*, vol. 46, no. 2, pp. 1–40, Nov. 2013.

[55] B. Ries, V. Interrante, M. Kaeding, and L. Anderson, "The effect of self-embodiment on distance perception in immersive virtual environments," in *Proc. ACM Symp. Virtual Reality Softw. Technol. VRST*, 2008, pp. 167–170.

[56] G. Rogez, M. Khademi, J. Supančič III, J. M. M. Montiel, and D. Ramanan, "3D hand pose detection in egocentric RGB-D images," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 356–371.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[58] V. Schwind, P. Knierim, C. Tasci, P. Franczak, N. Haas, and N. Henze, "'These are not my hands!': Effect of gender on the perception of avatar hands in virtual reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, May 2017, pp. 1577–1582.

[59] G. Serra, M. Camurri, L. Baraldi, M. Benedetti, and R. Cucchiara, "Hand segmentation for gesture recognition in EGO-vision," in *Proc. 3rd ACM Int. Workshop Interact. Multimedia Mobile Portable Devices - IMMPD*, 2013, pp. 31–36.

[60] M. Slater, M. Usoh, and A. Steed, "Depth of presence in virtual environments," *Presence, Teleoperators Virtual Environ.*, vol. 3, no. 2, pp. 130–144, Jan. 1994.

[61] Y. Tang, Z. Wang, J. Lu, J. Feng, and J. Zhou, "Multi-stream deep neural networks for RGB-D egocentric action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3001–3015, Oct. 2019.

[62] F. Tecchia, G. Avveduto, M. Carrozzino, R. Brondi, M. Bergamasco, and L. Alem, "[Poster] interacting with your own hands in a fully immersive MR system," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Sep. 2014, pp. 73–76.

[63] A. U. Khan and A. Borji, "Analysis of hand segmentation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4710–4719.

[64] A. Villegas, P. Pérez, and E. González-Sosa, "Towards a distributed reality: A multi-video approach to xR," in *Proc. 11th ACM Workshop Immersive Mixed Virtual Environ. Syst. MMVE*, 2019, pp. 1–4.

[65] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik, "The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 4, pp. 1643–1652, Apr. 2018.

[66] W. Wang, K. Yu, J. Hugonot, P. Fua, and M. Salzmann, "Recurrent U-Net for resource-constrained segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2142–2151.

[67] B. G. Witmer, C. J. Jerome, and M. J. Singer, "The factor structure of the presence questionnaire," *Presence, Teleoperators Virtual Environ.*, vol. 14, no. 3, pp. 298–312, Jun. 2005.

[68] W. Xiang, H. Mao, and V. Athitsos, "ThunderNet: A turbo unified network for real-time semantic segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1789–1796.

[69] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 9711–9717.

[70] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1038–1050, May 2018.

[71] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 175–185.

[72] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.

[73] Y. Zhu, K. Zhu, Q. Fu, X. Chen, H. Gong, and J. Yu, "Save: Shared augmented virtual environment for real-time mixed reality applications," in *Proc. 15th ACM SIGGRAPH Conf. Virtual-Reality Continuum Appl. Ind.*, vol. 1, 2016, pp. 13–21.

ESTER GONZALEZ-SOSA received the B.S. degree in computer science and the M.Sc. degree in electrical engineering from the Universidad de Las Palmas de Gran Canaria, in 2012 and 2014, respectively, and the Ph.D. degree from the Biometrics and Data Pattern Analytics (BiDA) Lab, Universidad Autónoma de Madrid, in June 2017. In October 2017, she joined the Distributed Reality Solutions Lab, Nokia Bell Labs, where she focuses on computer vision applied to mixed reality applications. She has carried out several research internships in worldwide leading groups in biometric recognition, such as TNO, EURECOM, or Rutgers University. Her research interests include biometrics with emphasis on face, body, soft biometrics and millimeter imaging, and computer vision techniques applied to egocentric perception, and mixed reality applications. She was a recipient of the Competitive Obra Social La CAIXA Scholarship, in 2012, the UNITECO Award from the Spanish Association of Electrical Engineers, in 2013, and the European Biometrics Research Award, in 2018.

PABLO PÉREZ received the Telecommunication Engineering degree (integrated B.Sc. and M.S.) and the Ph.D. degree in telecommunication engineering from the Universidad Politécnica de Madrid (UPM), Madrid, Spain, in 2004 and 2013, respectively. From 2004 to 2006, he was a Research Engineer with the Digital Platforms Television, Telefónica I+D. From 2006 to 2017, he has worked at the Research and Development Department, Video Business Unit, Alcatel-Lucent (later acquired by Nokia), serving as the Technical Lead of several video delivery products. Since 2017, he has been a Senior Researcher with the Distributed Reality Solutions Department, Nokia Bell Labs. His research interests include multimedia quality of experience, video transport networks, and immersive communication systems. He received the Doctoral Graduation Award for his Ph.D. degree.

RUBEN TOLOSANA received the M.Sc. degree in telecommunication engineering and the Ph.D. degree in computer and telecommunication engineering from the Universidad Autónoma de Madrid, in 2014 and 2019, respectively. In April 2014, he joined the Biometrics and Data Pattern Analytics (BiDA) Lab, Universidad Autónoma de Madrid, where he is currently collaborating as a Postdoctoral Researcher. His research interests include signal and image processing, pattern recognition, and machine learning, particularly in the areas of face manipulation, human–computer interaction, and biometrics. He is the author of several publications and also collaborates as a Reviewer in many different high-impact conferences (e.g., ICDAR, IJCB, ICB, BTAS, EUSIPCO, and so on) and journals (e.g., the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, ACM CSUR, and so on). He has participated in several National and European projects focused on the deployment of biometric security through the world. He has been granted with several awards, such as the FPU Research Fellowship from Spanish MECD, in 2015, and the European Biometrics Industry Award, in 2018.

REDOUANE KACHACH is currently a Researcher with the Department for Distributed Reality Solutions, Nokia Bell Labs, working mainly on the next generation of immersive video technologies. As professional background, he is a Senior Software Engineer with a large experience in software design and implementation of real-time, distributed, large-scale video processing, and distribution systems. In 2017, he joined Nokia Bell Labs as Research Engineer. He is working on the next generation of human communication solutions based on immersive video and mixed reality.

ALVARO VILLEGAS received the Telecommunication Engineering degree from the Universidad Politécnica de Madrid, Spain. He completed the M.B.A. Core Program at the ESCP Europe Business School. For the last 12 years, he has worked as an Innovation Lead in the field of digital video at Lucent, Alcatel-Lucent, and currently at Nokia, where he has filed more than 40 patents in the field. He is the Head of Nokia Bell Labs, Spain, the research center focused on the application of immersive media to human communications. His prior professional experience, always in the field of video innovation, was developed in Siemens, Telefónica Research and Development, ONO, Motorola, and Nagravision. He was awarded by Bell Labs with the Distinguished Member of Technical Staff Title.