HEADSET: Human Emotion Awareness under Partial Occlusions Multimodal DataSET

Fatemeh Ghorbani Lohesara, Davi Rabbouni Freitas, Christine Guillemot, Karen Eguiazarian, Sebastian Knorr

HEADSET: Human Emotion Awareness under Partial Occlusions Multimodal DataSET

Fatemeh Ghorbani Lohesara, Davi Rabbouni Freitas, Christine Guillemot, Karen Eguiazarian, and Sebastian Knorr

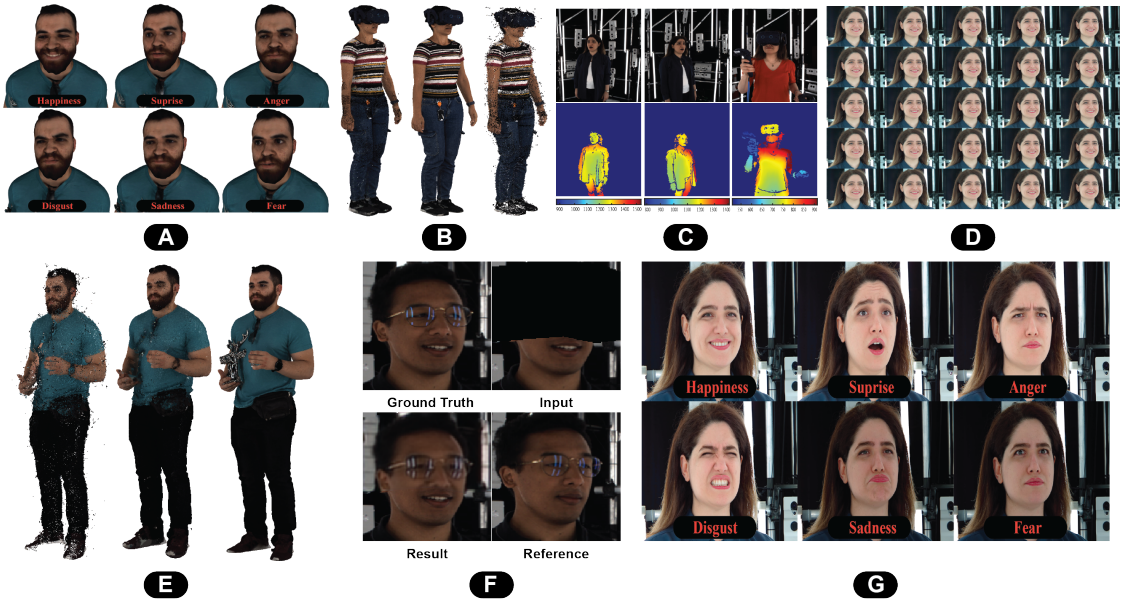

Fig. 1: Details of our proposed dataset (HEADSET) in terms of modalities and use cases. (A): Textured 3D meshes with 6 facial expressions, (B): three types of 3D point cloud with participant wearing an HMD, (C): RGB images and depth maps from 3 views, (D): multi-view representation of LF, (E): point clouds evaluation results, (F): result of HMD removal, and (G): classified LF images with 6 facial expressions.

Abstract—The volumetric representation of human interactions is one of the fundamental domains in the development of immersive media productions and telecommunication applications. Particularly in the context of the rapid advancement of Extended Reality (XR) applications, this volumetric data has proven to be an essential technology for future XR elaboration. In this work, we present a new multimodal database to help advance the development of immersive technologies. Our proposed database provides ethically compliant and diverse volumetric data, in particular 27 participants displaying posed facial expressions and subtle body movements while speaking, plus 11 participants wearing head-mounted displays (HMDs). The recording system consists of a volumetric capture (VoCap) studio, including 31 synchronized modules with 62 RGB cameras and 31 depth cameras. In addition to textured meshes, point clouds, and multi-view RGB-D data, we use one Lytro Illum camera for providing light field (LF) data simultaneously. Finally, we also provide an evaluation of our dataset employment with regard to the tasks of facial expression classification, HMDs removal, and point cloud reconstruction. The dataset can be helpful in the evaluation and performance testing of various XR algorithms, including but not limited to facial expression recognition and reconstruction, facial reenactment, and volumetric video. HEADSET and its all associated raw data and license agreement will be publicly available for research purposes.

Index Terms—Extended reality, multimodal dataset, virtual reality, volumetric video, light field

1 Introduction

- *Fatemeh Ghorbani Lohesara is with Communication Systems Group at Technische Universität Berlin. E-mail: ghorbani.lohesara@tu-berlin.de.*
- *Davi Rabbouni Freitas is with INRIA. E-mail: davi-rabbouni.de-carvalho-freitas@inria.fr.*
- *Christine Guillemot is with INRIA. E-mail: christine.guillemot@inria.fr.*
- *Karen Eguiazarian is with Computational Imaging Group at Tampere University. E-mail: karen.eguiazarian@tuni.fi.*
- *Sebastian Knorr is with Ernst-Abbe University of Applied Sciences Jena. E-mail: sebastian.knorr@eah-jena.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Nowadays, immersive visual technologies including Virtual Reality (VR), Augmented Reality (AR) and Mixed Reality (MR), or short Extended Reality (XR), play a key role in providing virtual experiences for users in various domains, such as XR teleconferencing, XR games, and XR experiences. Photorealistic representation of human interaction is essential for creating a life-like user experience and natural non-verbal communication among users. Specifically, the realistic representation of human facial expressions has a considerable impact on the quality of human interaction and communication [26].

An explicit volumetric representation can be visualized in XR applications by means of colored point clouds or textured meshes. Point clouds have drawn considerable interest due to their relatively simple process of collection and shortage of connectivity information. These features make them appropriate for real-time applications in XR, which require natural communication between users [41].

Nonetheless, the lack of human datasets, that are compulsory for volumetric representations, causes many challenges in developing photorealistic XR applications. The development of XR technologies and computer vision techniques is highly dependent on the quality of the datasets that can foster the progress of those fields and their evaluation. In contrast to the high number of research works in designing volumetric data generation algorithms [1,2,12,19,20,22,33,35,51] and three-dimensional (3D) reconstruction [37,38,43], there have not been enough studies in which we can discover relatively diverse and high-quality volumetric datasets with multiple modalities.

In this paper, we introduce a new multimodal database that consists of colored 3D point clouds, textured 3D meshes, light field (LF) images, and multi-view RGB-D images acquired from 27 ethnically diverse human participants. We have designed data collection tasks aimed at capturing all participants' posed and spontaneous facial expressions, and body language while speaking. Besides, we have also conducted an experiment for human full-body recording under occlusion caused by an HMD to collect human facial expressions under real-world XR scenarios from 11 participants out of 27. The database captured by our VoCap studio contains 31 ground truth depth cameras and 62 RGB cameras configured in 31 synchronized modules to also allow depth-from-stereo estimation. Additionally, we have used a Lytro Illum camera to collect the human facial expressions of all individuals from a frontal view to address the lack of publicly available LF face resources. Along with the raw and post-processed data, we also provide the labels of the facial expressions for the data collected from the VoCap studio (HEADSET-VoCap) and Lytro Illum LF camera (HEADSET-LF) as part of the main database.

The motivation to create such a dataset is to serve as the basis for research in XR-related use cases, especially in XR teleconferencing, where participants meet and interact in a virtual shared environment. In XR teleconferencing, participants are usually wearing headsets that need to be removed to enable eye contact. The knowledge about the emotion of the participant, who is wearing a headset, might increase the quality of the facial reconstruction. Many studies have focused on HMD removal [7, 29, 34, 54, 59], which is referred to as the task of virtual removal of HMD, which fill in the occluded color and geometric information of a user's face. The emergence of new MR glasses with emotion recognition capabilities and transparent displays, such as Meta Quest Pro and Apple Vision Pro, may increase the quality of the facial reconstruction results when removing the headset in such a study. HEADSET aims at providing ground-truth 3D models of individuals with and without wearing an HMD, and the HMD as an individual object captured with a volumetric capture studio. HEADSET can be utilized by further studies focused on reconstructing faces and gaze directions of the participants under the partial occlusions of the HMD. In this way, the data can be used to evaluate the person's identity after an occlusion removal algorithm is applied. Moreover, in XR teleconferencing scenarios, volumetric data needs to be compressed and streamed in high quality and low latency to increase the feeling of presence within an immersive environment. As the volumetric capture studio used in this study has the live-streaming capability, it allows XR teleconferencing in real time. On top of that, the proposed dataset has potential applications in rendering technologies, animation and simulation, perception, interaction, and user interfaces. We aim to contribute to the development of such approaches beyond their current capabilities to encourage a larger technical advancement in the fast-moving human-centric research domains.

This database can be used as a foundation for testing and validation of various computer vision problems such as 3D face and expression modeling, human activity and movement recognition, multi-view facial emotion recognition, facial reenactment, stereo matching, 3D compression, etc. on real-world high-quality data. In the design process of our data collection tasks and their contents, we have mainly focused on including HMD occlusions and human facial expressions during capturing to address issues related to XR communications. For example, HMDs significantly hinder the virtual experience as the headset covers the person's upper face and eyes. Our dataset thus includes representations of the individuals with and without wearing a VR headset, and the headset as an individual object for studies focused on reconstructing the faces of the participants under HMDs occlusions, as shown in Fig. 1. Research problems regarding the performance evaluation of facial expression recognition in multi-view RGB images and creating 3D models from a single image are also considered as the purpose of the usage of the proposed dataset. We have therefore incorporated these modalities as part of our dataset, an example of which is displayed in Fig. 1.

The main contributions of this paper can be summarized as follows.

1. We introduce a multimodal high-resolution database for immersive media productions in which LF images, RGB images, depth maps, textured meshes, and colored point clouds are crucial. HEADSET and its all associated raw data and license agreement will be publicly available for research purposes[1].

2. We collected the data taking into account the diversity of ethnicity and gender from our participants.

3. To the best of our knowledge, we are the first to provide volumetric data as a foundation for applications of emotion and face recognition under partial occlusions. This is done by capturing data of the individuals with and without an HMD to serve as ground truth for real-world XR scenarios.

4. Among many use cases, we selected three applications, in particular multi-view facial expression classification, HMDs removal, and visual quality assessment, for evaluating the dataset. The visual quality experiments are provided on different volumetric representations, i.e. textured meshes and point clouds.

The remainder of the paper is organized as follows. We first, review the related work in terms of available datasets in Sec. 2. Then, we present data acquisition steps and the capturing setup in Sec. 3. Participant selection criteria and ethical issues are also discussed in this section. Sec. 4 describes the data collection design in our user study. The data post-processing is explained in Sec. 5. We then report and discuss the results of three use cases of HEADSET. Finally, Sec. 7 summarizes our work.

2 Related Work

This section reviews related work on available human datasets for each modality, i.e. volumetric data, light field data, and RGB-D data with respect to human participants.

2.1 Volumetric dataset

CMUPanoptic [23] is the largest public volumetric dataset in terms of the number of capturing modules. In their work, human interactions of 8 participants in distinct social activities were recorded. The multi-view Panoptic Studio [23] consists of 31 HD, 480 VGA, and 10 RGB-D (Kinect v2) modules. Although CMUPanoptic stands as one of the largest currently available datasets, it does not provide hardware volumetric synchronization as the time alignment between the Kinect v2 RGBD streams is performed via a hardware modification. Zhixuan et al. released HUMBI [55], another publicly available and relatively large multi-view dataset. They captured the human body poses of 772 participants that participated in their study while wearing everyday clothes. The capturing setup included 107 synchronized HD cameras without any depth sensors. Therefore, no information about the ground truth geometry of the scene was acquired. Human4D [6] is another multimodal, marker-based approach to generate 4D data from volumetric sensors of 4 individuals performing 19 human daily activities. With the purpose of creating a dataset of high movement precision for the development of spatiotemporally aligned poses research, this dataset uses professional motion-capture (MoCap) markers and hardware synchronization for the multi-view data. However, this pursuit for high accuracy has its drawbacks: to produce more authentic movements, only 4 professional actors were selected to produce the 19 scenes, which is detrimental to

[1] https://webpages.tuni.fi/headset

the diversity aspect of the dataset. Also, the usage of MoCap apparel instead of natural clothing by the participants hinders the potential modeling of gaze, face, and body features.

Furthermore, one of the latest available dynamic point cloud datasets is CWIPC-SXR [41]. In this study, 23 human individuals performed activities in social XR settings, thus generating 45 unique scenes. However, the limited number of views used in the capture process – 7 Azure Kinect DK sensors – produced low complexity representations, resulting in low-resolution and non-watertight point clouds. In other words, the limited number of views during capture generated models with not so accurate textures and holes in the geometry from the occluded points.

The 8iVSLF dataset [25] is another dynamic voxelized point cloud dataset only containing 6 high-resolution single-frame models for 6 human individuals. The capturing setup included 39 synchronized RGB cameras configured in either 12 or 13 rigs.

Finally, Volograms & V-SENSE [36] have also introduced a small volumetric video dataset. The published dataset consists of three textured mesh sequences with differing characteristics and relatively short sessions. The dataset was captured in VoCap studios, which include 12 or 60 studio cameras.

In addition to the available volumetric datasets, HEADSET enables further research perspectives by collecting multimodal human data with the help of numerous camera modules. In our proposed dataset, along with the LF data, we present new volumetric sequences (HEADSET-VoCap) captured by a VoCap studio including 62 RGB and 31 depth cameras arranged in 31 synchronized camera modules. It contains post-processed textured meshes, point clouds with different resolutions, and multi-view RGB-D frames from 27 individuals displaying posed facial expressions and subtle body movements while speaking. Sequences under HMD occlusions are also part of the main database to introduce additional modalities compared to the available volumetric datasets. To better compare the proposed database with existing volumetric human databases as described before, we give an overview in Tab. 1.

2.2 Light field dataset

To the best of our knowledge, only four LF human face datasets have been made publicly available. The Light Field Face and Iris Database (LiFFID) [40] is the first human face dataset that contains images captured with an LF camera for the purpose of facial recognition. It comprises a group of 2D greyscale images created from the LF content captured by a Lytro lenslet camera. Nevertheless, LiFFID does not contain raw LF images, which is a considerable challenge for many research fields. The IST-EURECOM Light Field Face Database (LFFD) [48] is the second LF face dataset which includes both raw and rendered data from 100 persons, with 20 LF samples per participant acquired by a Lytro Illum lenslet camera in a controlled capturing setup with several facial variants. In [49], the authors introduced the Light Field Faces in the Wild (LFFW) and Light Field Face Constrained (LFFC) face datasets. LFFW includes 1908 LF images from 53 individuals captured in the wild in both indoor and outdoor environments without any predefined protocol. LFFC complements the LFFW dataset by including 1060 LFs from the same 53 participants acquired in constrained conditions. Despite the recent advances in LF face analysis [14] and facial expression recognition [49], highly accurate recognition results are still not achievable for some specific conditions due to the lack of data.

In our work, we present the HEADSET-LF dataset in addition to the volumetric data, which contains two subsets. The first one is collected from 27 participants showing 6 basic human emotions, totaling 162 LF images with corresponding labels for the facial expressions. The second one includes 10 LF frontal images of 10 individuals wearing a VR headset. Since this dataset will be publicly available, it may be used as the basis for the future validation and assessment of LF-based facial expression classification and recognition as well as facial reconstruction.

2.3 RGB-D dataset

Despite the wide availability of large RGB face image datasets [3, 27, 32], similarly sized datasets containing RGB-D face images are not available yet. RGB-D face datasets contain a limited number of samples [8, 31, 57], or they have been captured without considering HMD occlusions [60] and any additional modalities. Hence, researchers mostly tend to use a synthetic dataset with a high degree of variety in order to solve their research problems related to human faces. While the usage of a synthetic dataset reduces the potential of generalization to real-world data, this approach has been widely used in the literature. For example, for HMD removal/ facial reconstruction, the authors of [34] built a data synthesis pipeline to create a synthetic dataset of RGB-D images with a random pose, ambient illumination, and expression of faces based on the Basel Face Model (BFM) 2017 [15]. To address the aforementioned challenges, our dataset contains additional sequences under HMD occlusions in order to be used as the testing set for the future validation and assessment of such research work. Along with LF and RGB-D data, HEADSET also includes 3D point clouds and 3D textured meshes with an average number of frames of 272 for every 27 individuals. The sequences under HMD occlusions also contain 58 frames per participant for 11 participants out of 27.

3 Data Acquisition

In this section, we provide more details of the data acquisition process with regard to the capturing setups, VoCap studio and Lytro Illum camera, and the performed steps for participant selection.

3.1 Capturing Setup

3.1.1 Volumetric capture studio

We have utilized a 3D VoCap Studio (Mantis Vision Volumetric Capture System[2], version: studio ring) for capturing the volumetric dataset. A custom room setting with a cylindric recording area was employed (radius: 1.6 m, height: 2.5 m). Similarly, the capture rigs were placed in a cylinder with radius of 2.5 m and height of 3 m. The VoCap studio contains an aluminum frame with a black background and adjustable lighting. Hence, the capturing scene was illuminated by 34× Quasar Science Q50XG lights. They were spread evenly around the studio in order to provide enough light for our capturing scenes. The floor of the recording space is also black. The black backgrounds and floor reduce reflections during the capturing process. The complete setup is shown in Fig. 2. The studio has three types of uEye cameras and 31 camera modules. Each camera module has a laser and monochromatic UI-3140xCP-M for structured light based depth estimation. In addition, it has two UI-3080xCP-C or UI-3280xCP-C cameras for color information. Therefore, a total of 62 RGB cameras (2054×2456 pixel) and 31 depth cameras (1024×1280 pixel) were used.

The depth camera supports several modes by which the frame rate, resolution, exposure time, operating range of the module, and region of interest can be modified. The modes of the color camera, including frame rate, resolution, field of view, aspect ratio, and format can be also determined. In our work, all the raw data had been captured with 25 frames per second (fps) during the data collection.

The calibration and synchronization of the VoCap studio was carried out once before starting the capture, and the calibration parameters are enclosed within each dataset based on its modality. As the VoCap studio can be seen as a single fully calibrated capturing device for capturing high-quality 3D models, the recorded volumetric data can be used as ground truth. However, we extracted 3 RGB-D module outputs separately based on their field of view and our experiments' analysis. Thus, the exact internal and external camera parameters have also been provided within the RGB-D dataset. When the participant stands in the center, the distance to the cameras varies roughly from 80 cm to 120 cm. While standing in the center and looking ahead, 11-13 camera modules have a good view of the person's face within about 130 degrees angle in the individual's field of view.

[2] https://mantis-vision.com/3d-studio-3iosk/

Table 1: Comparison of state-of-the-art volumetric datasets with ours in terms of modalities and capturing details.

Dataset	Participants	Description	Data type	Capturing modules
CMUPanoptic [23]	8	Participants performing social activities	Multi-view RGB-D, 3D point clouds	31 HD, 480 VGA, and 10 RGB-D
HUMBI [55]	772	Human body expressions	Multi-view RGB, 3D meshes	107 HD
Human4D [6]	4	Professional actors performing full-body movements and expressions	Multi-view RGB-D, point clouds, 3D meshes	24 motion capture and 4 depth sensors
CWIPC-SXR [41]	23	Human interaction in social XR settings	Multi-view RGB-D, 3D point clouds	7 Azure Kinect DK sensors
8iVSLF [25]	6	Full-body of a human participant	3D point clouds	39 RGB cameras in 12/13 rigs
Volograms&V-SENSE [36]	3	Monologue and dancing	Meshes with texture images	12/60 camera studio
HEADSET	27	Posed facial expressions and subtle body movements w/o VR headset	Multi-view RGB-D, 3D point clouds, 3D textured meshes, light field	62 RGB, 31 depth, and 1 Lytro Illum cameras

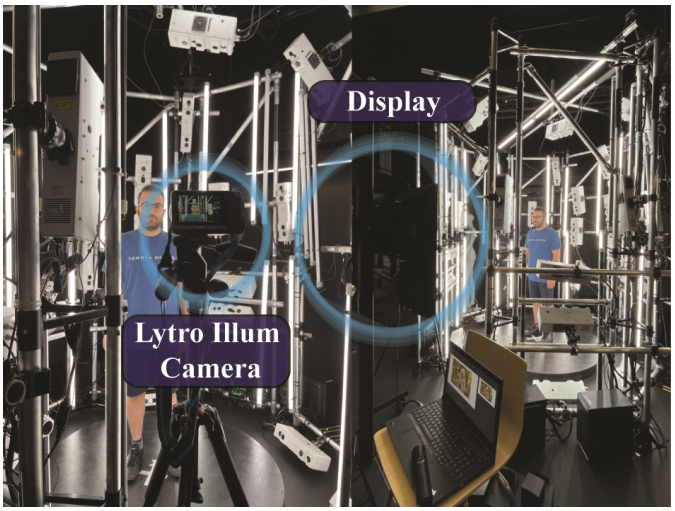

Fig. 2: The complete capturing setup for the data collection. In addition to the VoCap studio, we also used one Lytro Illum camera, one display, and a microphone.

3.1.2 Lytro Illum light field camera

Complementary to the VoCap studio, we added one microphone to record the audio signal, one frontal Lytro Illum LF camera to capture the frontal view, and one large display for showing the necessary content of the experiment to the participant. The Lytro Illum is a 40 megaray LF camera with constant F/2.0, 8X optical zoom, and a 4-inch tilt screen. The camera provides data in Light Field Raw (LFR) and depth map (PNG and TIFF, 2022×1404 pixel) file formats, which are contained in our dataset. The LFR file stores the image information in an uncompressed lenslet image before further processing. The relative depth of field coordinates (λ_{max} and λ_{min}) along with the calibration information have been also included in the dataset, which has been extracted using the Lytro Desktop Application. The camera height and zoom ratio were properly adjusted to capture the individual's face based on its height. During the capturing process, the camera was located in front of the participant while the person was standing at the VoCap studio's central point (153 cm distance from the Lytro camera's location). Calibration was performed before every recording session, using a 3 m, 16 mm steel tape measure. The complete setup and the location of the display, and LF camera are depicted in Fig. 2.

3.2 Participant selection

3.2.1 Ethical issues

Before taking part in the data collection tasks, each individual had to read and sign an information sheet and consent form, which allowed the use of data for research purposes and data publication.

The participants were also supposed to read the safety training material. All of the possible risks of taking part in our study, such as physical discomfort, which may potentially happen due to wearing a VR headset or posing facial expressions, have been mentioned in the participant information sheet. In order to collect high-fidelity records, the volunteer was asked to avoid large head and body movements during the capturing. The ethical issues in our work had been carefully considered and were fully approved by the Academic Ethics Committee of the Tampere Region, Tampere University. The complementary explanations of the ethics procedure in our work can be found in the supplementary material.

3.2.2 Participants

We looked for people who were interested to take part as volunteers in our study. There were no specific criteria for participant inclusion, with the exception of each participant having to be above the age of 18 years. Although we did not inquire about the participants' ethnicity due to ethical concerns, we attempted to keep the dataset ethnically diverse to the best of our efforts. There are 19 male participants with an average age of 26.37 years and an average height of 178.05 cm. The corresponding figures for the 8 female participants are 27.5 years and 165 cm. 10 out of 27 participants were wearing glasses at the time of the capture, there being 9 male and 1 female.

4 Data Collection Design

Our data collection tasks aimed to capture posed expressions (task A), spontaneous facial expressions, and body poses while speaking (task B). Moreover, we have also collected human face and full-body recording under occlusion caused with a VR headset (task C), i.e. we have designed three types of assignments for data collection.

Before starting the data collection, two training examples were given to the volunteers to explain what they were supposed to do during each session: one before A and B, and the other before C. Thus, tasks A and B were carried out uninterrupted. In the first training session, we explained the tasks of tasks A and B to the participants and showed them sample images (dissimilar to the images shown in the effective tasks but in the same category). In the second training session, which was performed after finishing assignments A and B, we asked the participant to wear a VR headset. More details of the data collection task are explained in the following subsections.

4.1 Data collection Task A

This data collection task consisted of showing the volunteer 6 basic human emotions on a big display screen in front of the person and the VoCap studio. The basic emotions included *happiness, surprise, anger, disgust, sadness,* and *fear*. The target emotions were defined and displayed as described in [53]. Presenting facial images on a display was our main way to elicit such expressions. The participants were asked to look at each picture that appeared on the display and then to try to mimic the expression that had the same semantic meaning as the displayed one.

4.2 Data collection Task B

In this assignment, we displayed three pictures to the volunteer with background sounds related to the content of the pictures. These three images contained animals, a baby, and a nature scene, respectively. The participants were then asked to look at each picture and describe it and their feelings about each of them in their own words. The spoken language was either English or the participant's mother tongue based on their own preference. We encouraged them to express their thoughts in their mother tongue so that they could generate facial expressions that were as close to natural as possible when speaking. In total, 15 out of 27 participants spoke in English, and the rest preferred to speak in their mother tongue.

4.3 Data collection Task C

In the final task, the participants were asked to wear an HTC Vive Pro Eye headset [21] while standing in the center of the capturing studio, where they looked at different points on the headset's display. The participants were supposed to move their body around without changing their location. During this process, participants mostly showed Neutral emotions while looking at different cameras. In addition to recording sequences from the volunteers, we also reconstructed a 3D model of the VR headset, which we used for task C, for 2 frames. This recording aims at providing a ground-truth 3D model of the occlusion object for further studies focused on reconstructing faces and gaze directions of the participants under the partial occlusions of the headset. In cases where identity preservation is critical, such as HMD removal and facial expression reconstruction, we have provided representations of the individual wearing a VR headset and without it, and the headset as an individual object. In this way, all three types of data can be used to evaluate the identity of the person after an occlusion removal algorithm is applied.

5 Data post-processing

We recorded over 5 hours of raw data with 25 frames per second (fps). However, it was neither feasible nor necessary to post-process all of them mainly because of the high computational cost and memory issues. Therefore, we post-processed each of the sequences based on data type usage with different segments. The post-processing frame rate was 3 fps for data collection tasks A, B, and C. It is worth mentioning that in addition to the post-proceed data, the raw captured data @25 fps and camera calibration parameters are also made available. Each data type has been organized according to the participant's identification number and frame number. Detailed instructions on data structure and synchronization information are also given within each dataset. In this section, we go through each data type and explain the applied post-processing steps.

5.1 Textured 3D meshes

In order to reduce the amount of data processing as well as to avoid collecting too many similar frames, we decided to reduce the sampling rate of the post-processed data. In order to generate the 3D meshes, the Poisson surface reconstruction [24]technique was applied from the raw point cloud data. In addition to capturing sequences from the participants, we also built a 3D model of the VR headset, which we used for data collection task C, for 2 frames. Fig. 3 illustrates examples of reconstructed textured meshes post-processed after recording, and RGB images captured during the recording of data collection tasks B and C.

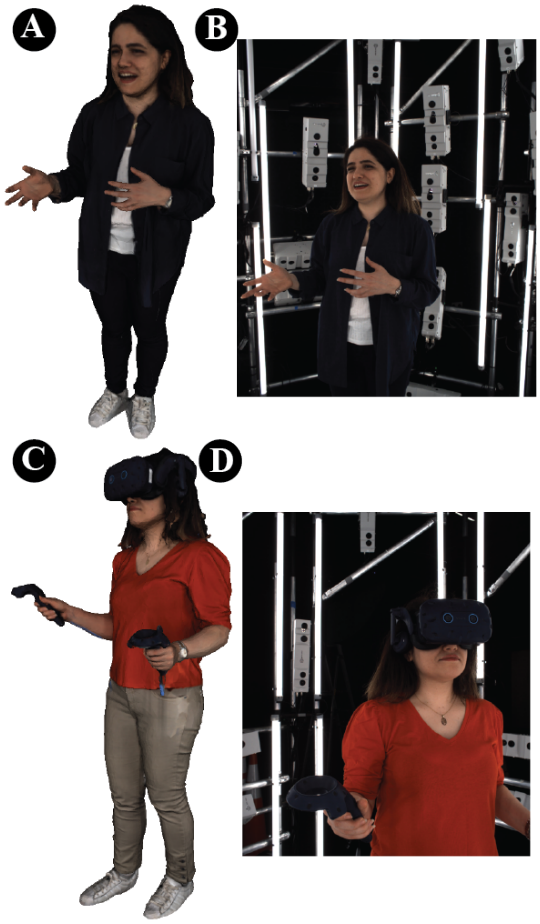

Fig. 3: Example of reconstructed textured meshes. (A): full-body 3D model of a participant, (B): RGB image captured by camera number 30, (C): full-body 3D model with HMD occlusion, and (D): RGB image captured by camera number 16.

5.2 Colored 3D point clouds

The raw point cloud data is obtained by generating the geometry from the ground truth depth maps captured by the 31 depth cameras of the VoCap Studio. The 2 stereo images from each capture module further improve the geometry by applying depth-from-stereo, while also coloring the scene's points. Fig. 4 illustrates examples of reconstructed point clouds acquired after recording and corresponding RGB images captured from an individual wearing glasses during the capturing process.

Due to the sparse and noisy nature of the raw point clouds (Fig. 4-B) – which contain around $\sim 300,000$ points –, we also provide post-processed versions of them. This is done by removing outlier points and applying a Poisson surface reconstruction [24], as done in [16], to increase their resolution. Afterwards, we sample the points from the mesh [10] with a surface density – the number of points per square unit – of 0.05, resulting in point clouds of around $\sim 900,000$ points. One example is depicted in Fig. 4-C. Finally, even though the post-processed point clouds increase the models' resolution, certain materials like extremely non-Lambertian objects, *e.g.* mirror-like surfaces, are not well represented only from the RGB-D images due to the lack of geometry of the raw depth data. Thus, we also provide an additional type of post-processed point cloud, which is sampled from textured meshes. An example of this type of a post-processed point cloud is illustrated in Fig. 4-D.

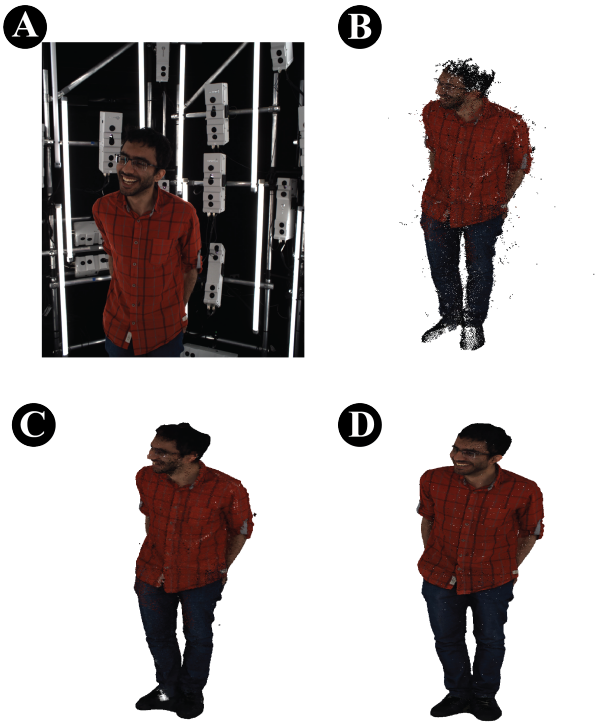

Fig. 4: Example of colored point clouds of a participant wearing glasses in task A. (A): RGB image captured by camera number 30, point cloud representation from (B): raw data, (C): post-processed, and (D): sampled from textured meshes.

5.3 RGB-D

We have also collected RGB-D images for tasks A and B from two of the capture modules based on the cameras' field of view and the capture setup's layout. Their corresponding indices are 1 and 30. These indexes were chosen since they provided a good field of view for capturing the volunteer's facial expressions. The distance between the depth cameras and between the RGB cameras of modules number 30 and 1 is 989.565 mm and 991.126 mm, respectively. At the beginning of each sequence, the participants were trying to understand the first task. To that end, we decided to check all the sequences manually and remove the redundant frames from the start and end of capturing to avoid the collection of many similar and incomplete frames. The script for exporting the depth maps for each frame is also published together with the RGB-D dataset. The script that reads, processes, and visualizes the camera transformations is included in the dataset along with the exact positions of each camera. Therefore, each subset of the RGB-D data processed for all data collection tasks contains the extrinsic and intrinsic calibration matrices for both RGB cameras and the extrinsic matrix for each module's depth sensor at the time of capture.

For task C in which the person was wearing a VR headset, RGB-D images have been collected from one frontal module (number 16) for 20 seconds at 3 fps. Here, module number 16 provided the frontal view because the participant was looking in the opposite direction compared to data collection tasks A and B.

In Fig. 5, a sample of RGB images and depth maps from camera number 1 and 30 is shown, where the individual was performing the "Surprise" expression. An example of RGB-D representation from module 16 in task C, in which the volunteer is wearing an HMD, is also depicted in Fig. 5 (C, F).

5.4 Light field

The LF image dataset consists of two subsets according to their content. The first one was collected from 27 volunteers showing 6 basic human emotions, totaling 162 LF images. The data is labeled into 6 classes based on the participant's performed emotion. The second one includes

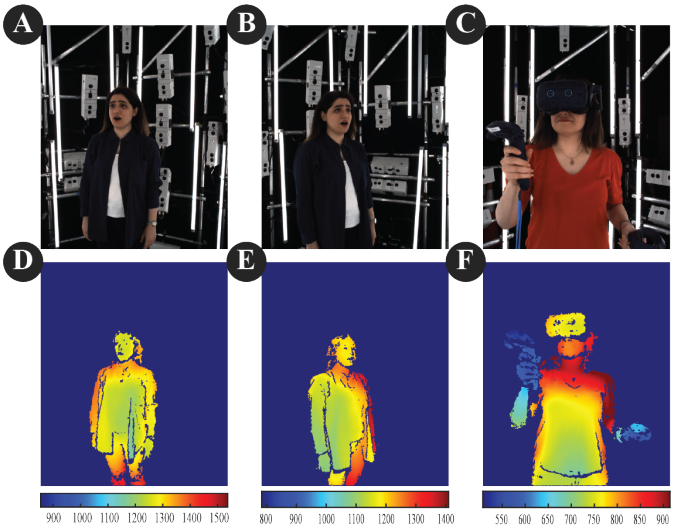

Fig. 5: Sample of RGB images and depth maps from three views. (A,D): RGB-D image of module number 30, (B,E): RGB-D image of module number 1, (C,F): RGB-D image of module number 16.

10 LF frontal images of 10 people wearing an HMD. Along with LF raw data (.lfr files), their related depth maps are also available in TIFF and PNG file formats.

Tab. 2 summarizes the details of the HEADSET multimodal dataset captured by the VoCap studio and Lytro Illum camera.

5.5 Post-processed labeled data

We have also created two RGB labeled subsets of our main database which include human facial expressions. The first one is the multi-view representation of LF data captured by the Lytro Illum camera, and the second one contains RGB images from two non-frontal views captured by the VoCap studio. The label of each image has been defined based on the ground truth emotion described in task A. The images that we used further for the evaluation of the facial expressions classification (FEC) in our dataset include VoCap RGB data (HEADSET-VoCap), and multi-view representation of light field data (HEADSET-LF), both in PNG file format. HEADSET-LF is created from sub-aperture images of the LF raw data as multi-view RGB images. For this work, each LF raw data is converted into a 5×5 RGB view matrix. It is noteworthy that in some cases the participant showed "Neutral" emotion instead of the required emotion. Thus, we removed the samples that are apparently not matched to the ground truth label by human observation. However, we made both the original and the modified datasets with labels available. The number of RGB images in the modified datasets, which we used for the evaluation, are as follows: HEADSET-VoCap: *{Anger: 363, Disgust: 266, Fear: 209, Happiness: 264, Sadness: 284, Surprise: 420}* and HEADSET-LF: *{Anger: 650, Disgust: 550, Fear: 450, Happiness: 675, Sadness: 375, Surprise: 575}*, which are acquired from the multi-view RGB representations of the raw LF images.

For our evaluations and experiments, which we describe in Sec. 6, we first applied a deep cascaded multi-task framework method for face detection (MTCNN) proposed in [58] on both labeled datasets in order to make the evaluation faster. We then checked the facial images and landmarks of all views of the dataset and proved that the facial region is detected in all the frames.

Fig. 6 depicts two non-frontal views (left: camera number 1, and right: camera number 30 of the VoCap studio) of HEADSET-VoCap, and one frontal view (middle: captured by the Lytro Illum camera, central sub-aperture) of HEADSET-LF as examples for the "Happiness" class after applying the face detection algorithm.

6 Experimental Results and Dataset Evaluation

We conducted multiple experiments using different types of our data in order to report HEADSET's performance compared to similar ones in

Table 2: HEADSET modality details with regard to its content. All the raw data captured by the VoCap studio @25fps is additionally available in .mvx file format with a total size of 2598 GB.

Data type	Participants (out of 27)	Content	Occlusion	Capturing modules	Avg. # of frames	Avg. size	Format
Colored point clouds	27	6 posed expressions and subtle body movements	Natural (caused by glasses or hair)	31	272	4.33 GB	.ply
Colored point clouds	11	Subtle body movements	HMD	31	58	885 MB	.ply
Meshes with textures	27	6 posed expressions and subtle body movements	Natural	31	272	2.07 GB	.obj
Meshes with textures	11	Subtle body movements	HMD	31	58	339 MB	.obj
RGB-D	27	Posed and spontaneous facial expressions	Natural	2	455	6.7 GB	.png
RGB-D	11	Subtle facial movements	HMD	1	20	595 MB	.png
LF	27	6 posed expressions	Natural	1	6	377 MB	.lfr
LF	10	Face	HMD	1	1	65 MB	.lfr

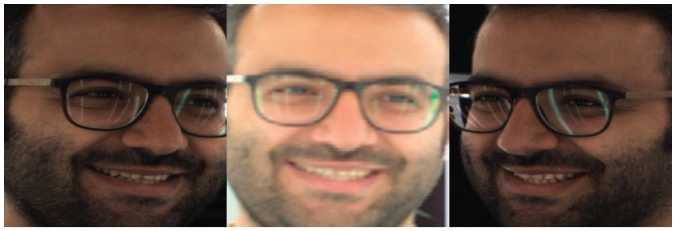

Fig. 6: Three synchronized views (two non-frontal views in HEADSET-VoCap, and one frontal view in HEADSET-LF) of detected faces showing a "Happiness" expression.

a multitude of applications. We first present the volumetric assessment of sequences under the HMD occlusion compared to a similar currently available dataset. Then, we evaluate the dynamic scenes in the context of compression with two state-of-the-art 3D codecs for voxelized point clouds. Afterward, we focus on two popular computer vision problems which can be a use case of the proposed dataset. The first application involves facial expressions classification in HEADSET-VoCap and HEADSET-LF collected from Experiment A, as described in Sec. 5.5 in order to prove the expression variations in the collected data. We also present the results of a deep video inpainting model on our dataset for solving the HMDs removal problem as the second application. In this section, we explain each of the experiments in detail.

6.1 No-reference volumetric assessment of headset-wearing participants

Although the volumetric datasets from Tab. 1 provide different scenes for immersive applications, including typical social XR situations as well as body poses and movements, few of them include a photorealistic representation of the participants' occluded expressions. More precisely, only the CWIPC-SXR dataset [41] contains such data from the aforementioned datasets, where two scenes depict three individuals performing actions while wearing an HMD. Our HEADSET data contains volumetric information (both meshes and point clouds) from high-resolution captures for 11 different persons wearing an HMD, while also providing scenes of the same individuals without it.

In order to assess the quality of our generated volumetric data, we estimate the subjective quality of our 3D data using a no-reference (NR) metric to evaluate the post-processed point clouds derived from the textured meshes as described in Sec. 5.2. The usage here of an NR point cloud quality assessment (PCQA) is paramount due to the lack of a reference point cloud to compare to. That is, potential distortions over the geometry and texture data are due to the nature of the capturing and processing pipeline of the dataset to generate the scenes. Therefore, it is important for the selected NR-PCQA metric to have a good generalization capability over the different kinds of possible distortions regarding the geometry and the attributes.

Table 3: Average Pseudo Mean Opinion Scores of participants wearing a head-mounted display. Higher is better. Results were averaged from 20 frames of 3 individuals from [41] and 11 individuals from our dataset.

Point cloud type	MOS ↑
CWIPC-SXR [41]	2.885
Raw	4.127
Post-processed	4.443
Sampled from texture meshes	**4.853**

With that in mind, we assess our generated data for participants wearing an HMD via the ResSCNN NR metric [28], which leverages a sparse convolutional neural network to estimate the subjective quality of point clouds without the usage of reference models. To properly evaluate our and the CWIPC-SXR scenes, we use a ResSCNN model trained on a large-scale point cloud quality assessment dataset, which contains 104 reference point clouds with more than 22,000 example cases with 31 different types of distortion over the geometry and the attributes data. These distorted samples are annotated with a pseudo mean opinion score (MOS) to subjectively evaluate the 3D data (see [28] for a more detailed explanation). In short, this pseudo MOS is a scale of five quality levels in the range $[1,5]$, where 1 means that the distortions significantly hinder the perception of the scene and 5 means that almost no distortion is perceived.

Our experiments are performed over 20 frames for each of the scenes. The 11 participants of HEADSET-VoCap are evaluated against the two scenes from the CWIPC-SXR [41] dataset that contains participants wearing an HMD: scene 14 ("Rock-paper-scissors in VR"), containing two persons, and scene 19 ("Boxer in VR"), with one. As recommended in [28], all the sequences' coordinates were scaled in the range of $[0-2000]$ for the evaluation.

Results from Tab. 3 suggest a superior subjective quality of our scenes than the ones in [41] according to their pseudo-MOS. In particular, these results show that even our raw types of point clouds still present a decent subjective quality, even in the presence of outliers. Our post-processed scenes, which include the outlier removal in the post-processing step – a procedure that is also done for the data in [41] –, show an even greater improvement over our raw types, which is expected, with the ones derived from the textured meshes performing the best out of them. Moreover, we not only provide a larger number of

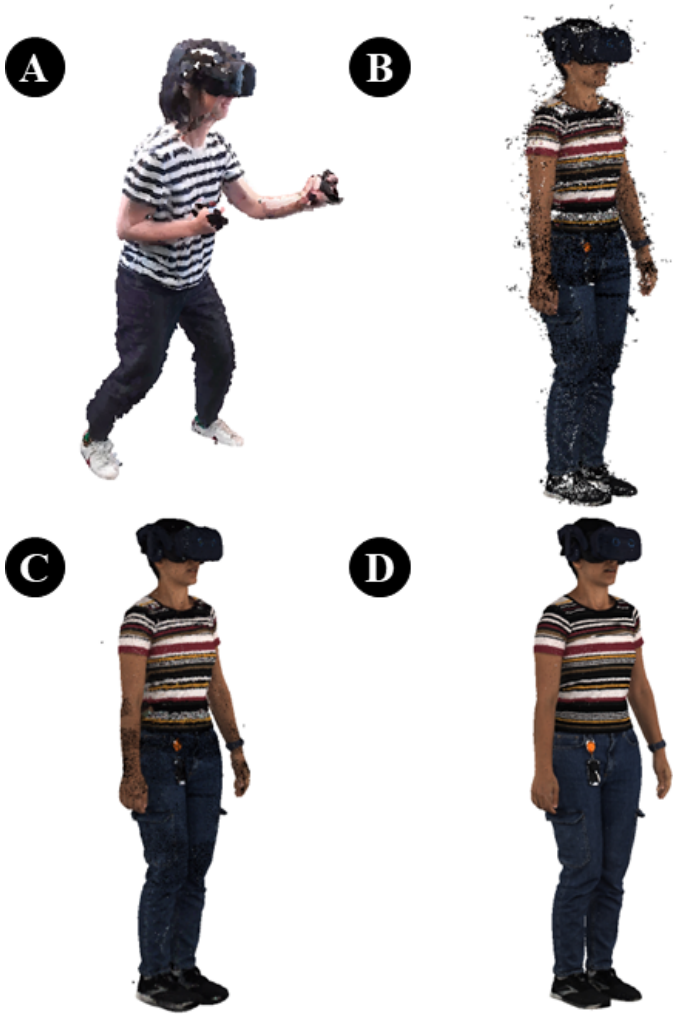

Fig. 7: Visualization of a frame from a) “Boxer in VR", from [41] and versions for our three types of point clouds: b) raw, c) post-processed, and d) derived from textured meshes, for Participant 19 of our dataset.

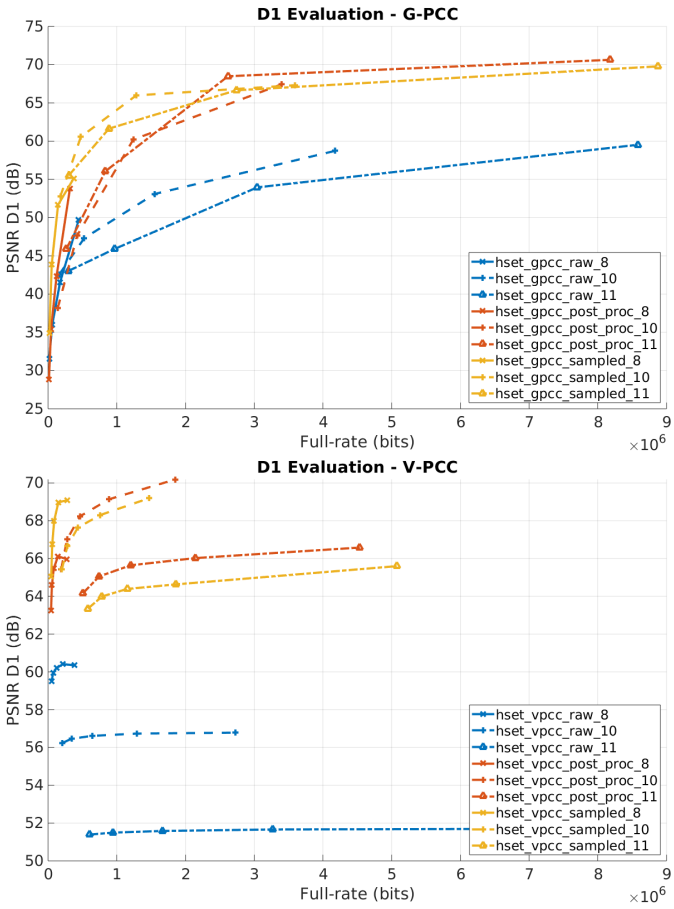

Fig. 8: RD Results for the average D1 metric from 4 frames for codecs G-PCC (top) and V-PCC (bottom).

participants, but also their ground truth non-occluded physiognomies, making our data suitable for applications targeting the study of facial occlusion directly over the 3D data.

Fig. 7 also shows an illustration of four 3D models of the compared datasets, with one frame from the sequence “Boxer in VR" from [41] and the other three being our three types of point clouds. Notice how the sequence from [41] in Fig. 7-a) presents significant distortions in the geometry from “holes", *i.e.* missing points, and also in its colors. On the other hand, the frame taken from the raw point cloud for participant 19 of our dataset in Fig. 7-b) appears to have less evident distortions for the colors, while presenting a more significant number of outlier points. This number of outliers is greatly reduced for our post-processed sequence in Fig. 7-c), although it presents some minor holes and lacks some of the finer texture and geometry details. Our 3D model derived from the textured meshes, which can be seen in Fig. 7-d), not only provides a watertight geometry but is able to depict some of the scene’s finer details, such as the watch and the harness on the person’s belt.

6.2 Point clouds evaluation in terms of compression

In order to popularize applications that provide virtual experiences with low latency, like XR telepresence, XR games, and free-viewpoint videos, it is paramount for the volumetric data to be efficiently conveyed in real-time. As such, we benchmark our dynamic scenes in the context of compression with two state-of-the-art 3D codecs for voxelized point clouds from the Moving Picture Experts Group (MPEG): the geometry-based point cloud compression (G-PCC) and the video-based point cloud compression (V-PCC) standards [47]. Since our dataset consists of sequences with multiple frames of volumetric data for each scene, we provide a testbench for the development of both static – or intra-frame – and dynamic – or inter-frame – methods for volumetric video compression. Hence, we selected V-PCC due to its suitability for temporal video compression, and G-PCC for its profile of static data compression. More details on both solutions can be found in [18,47].

As both codecs require the point cloud data to be voxelized, that is, the points are quantized into volumetric elements, we constrain our data into three different voxel grid resolutions, $b = \{8, 10, 11\}$. This quantization procedure allows us to assess the visual quality and size of our data with regard to different resolution levels, and how it manifests into applications where the data has to be conveyed. Therefore, we assess our three different types of point clouds, as outlined in Sec. 5.2, with a rate-distortion (RD)-based approach in order to evaluate three key outcomes: 1) evaluate how the different characteristics (number of points, density, "watertightness", etc.) from these three types are materialized in a transmission context; 2) observe which voxelized resolutions prove to be more adequate in a live-streaming case according to the data that we generated; 3) assess the conveyance of our data both in an inter-frame and an intra-frame scenario.

To address objective number 3), 4 consecutive frames of each of our 27 scenes were selected to be conveyed through V-PCC and G-PCC. Further information about the selection of the frames from the scenes and the specifications used for both codecs are explained in the supplementary material.

The naming convention for our results is based on the type of the point cloud, the codec used, and the resolution that was applied to the data, with the format *hset-codec-*pc_{type}*-b*, where $pc_{type} = \{raw,$

post_proc, *sampled*}, $codec = \{gpcc, vpcc\}$ and $b = \{8, 10, 11\}$. The RD results are constructed such that the rate consists of the total rate, in bits, required for transmission. For the quality metric, we evaluate the geometry by using the point-to-point metric, also known as D1 [52], and for the attributes, we use the Peak-Signal-to-Noise-Ratio (PSNR) for the Y, U, and V channels of the original and decoded point clouds.

The results for the average D1 metric from 4 frames of our 27 sequences can be observed in Fig. 8, while the results for the attributes are presented and discussed in the supplementary material. Notice that, even though the bitrate increase is expected and increases with the resolution, the conveyed size increases more significantly when going from a resolution of 10 to 11 bits, which is particularly noticeable for the *hset-vpcc-raw* point clouds due to their lower density in comparison to *hset-vpcc-post_proc* and *hset-vpcc-sampled*. The significant decrease in the D1 metric also indicates that the number of decoded points is much larger than the original voxelized point clouds. You can observe this effect for b = 10 for the *hset-vpcc-raw* and b = 11 for *hset-vpcc-post_proc* and *hset-sampled*, with the same happening for G-PCC in a lesser degree. Finally, note that V-PCC provides a better RD performance over G-PCC for the 4 transmitted frames, in particular for $b = 10$. This is to be expected, due to V-PCC's inter-frame scope.

6.3 Facial expression classification (FEC)

As described, the MTCNN detector without any margins is utilized before applying the FEC models, so that most parts of the background such as hair follicles are not present. As a result, the learned facial features are more suitable for emotional analysis.

To make comparisons for evaluation in FEC, we selected four similar datasets namely JAFFE [30], AffectNet [32], AFEW [11], and VGAF [50], which are used as benchmark by prior works on the FEC problem. In the following, we briefly explain each dataset's characteristics with regard to its content.

- *JAFFE* [30]: The Japanese Female Facial Expression (JAFFE) database contains 213 grayscale images of acted Japanese female facial emotions. All the images are resized into (256×256 pixel). It includes 7 basic human facial emotions (*Anger, Disgust, Fear, Happiness, Sadness, Surprise,* and *Neutral*). For the comparison, we used all 213 images as the testing set.

- *AffectNet* [32]: This RGB image dataset includes 287,651 images in its training set, and 3,999 images in its validation set of 8 human facial expressions (*Anger, Disgust, Fear, Happiness, Sadness, Surprise, Contempt,* and *Neutral*). We used 7 classes (excluding *Contempt* emotion) of the original validation set of 3,499 images for testing purposes of AffectNet. The faces are detected by the authors of the dataset before evaluation.

- *AFEW* [11]: The AFEW dataset with 773 train and 383 validation samples contains audio-video short clips acquired from TV serials and movies with different poses, spontaneous expressions, and illuminations. They are grouped by a single emotion label to the video clip from 6 basic emotions (*Anger, Disgust, Fear, Happiness, Sadness,* and *Surprise*) and *Neutral*, as described in [46].

- *VGAF* [50]: This dataset shows a wide amount of variations in both training and validation sets. The data contains 2,661 clips and 766 videos for training and validation, respectively. The FEC problem in VGAF is to classify each video into 3 classes (*Positive, Neutral,* and *Negative* emotions) [46].

Then, we applied five state-of-the-art methods for facial expression classification on the aforementioned and our HEADSET dataset. These methods include the models Ad-Corre (trained on RAF-DB [27] dataset for 7 classes) [13], ResMaskingNet (trained on FER-2013 dataset [17] for 7 classes) [39], as well as lightweight EfficientNet-B0 (trained for 8 classes), EfficientNet-B2 (trained for 8 classes), and EfficientNet-B2 (trained for 7 classes) [44–46], which were trained on the VGGFace2 dataset [4]. The accuracy performance metric is then computed for all datasets. Tab. 4 gives a summary of the accuracy measures for all five models on our datasets as well as on the validation sets of AFEW [11], AffectNet [32], VGAF [50], and all images of the JAFFE dataset [30]. As EfficientNet-B0 and EfficientNet-B2 are capable of extracting emotional features in video frames [46], AFEW and VGAF datasets have only been used for video-based emotion recognition. The details of F1 scores are presented in the supplementary material due to space limitations.

It is worth mentioning that the usage of a model trained on 8 or 7 classes to predict 7 or 6 emotional categories presents a slightly lower accuracy, though it is more general as it can be used to predict either 8, 7, or 6 emotions [44].

As proved in [9], multi-view representation of light field images recorded by a Lytro Illum camera can provide complementary information beneficial for face recognition. Thus, we have processed the raw data of Lytro Illum camera to render LF images as multi-view RGB images, each one collected from a slightly shifted point of view. For this experiment, each data is transformed in a 5x5 RGB view matrix, each view with size 620×432 pixels. As it is observable in Tab. 4, the results of all five models on our datasets are comparable to other benchmark datasets in FEC.

6.4 HMD removal

For facial reconstruction of the areas occluded by an HMD in the collected video frames, we used a GAN-based method for deep video inpainting, named Learnable Gated Temporal Shift Module (LGTSM) introduced in [5], with the same hyperparameters and training procedure. The following modifications were applied. We added an additional self-attention layer and its non-local operations, introduced in [56], in its encoder part before the dilated convolution layers in order to capture long-range dependencies between different regions of an input feature map. Specifically, the self-attention module takes a feature map as input and applies three convolution layers to compute the key, query, and value vectors. The kernel size and stride for each convolutional layer are set to 1×1 and 1, respectively, to capture spatial relationships between neighboring feature map locations while keeping computational costs low.

We used FaceForensics dataset [42] and our collected RGB data from VoCap as the training set and validation set, respectively. FaceForensics is comprised of 1,004 videos including more than 500,000 frames with faces collected from Youtube. They consist of only frontal faces cropped to 128×128. The whole RGB video sequences in our dataset collected from VoCap include 11,584 frames captured by camera number 30 and 1 from all 27 volunteers, as described in Sec. 5.3. We first applied MTCNN for face recognition, then resized the frames to 128×128 to make them similar to the training set.

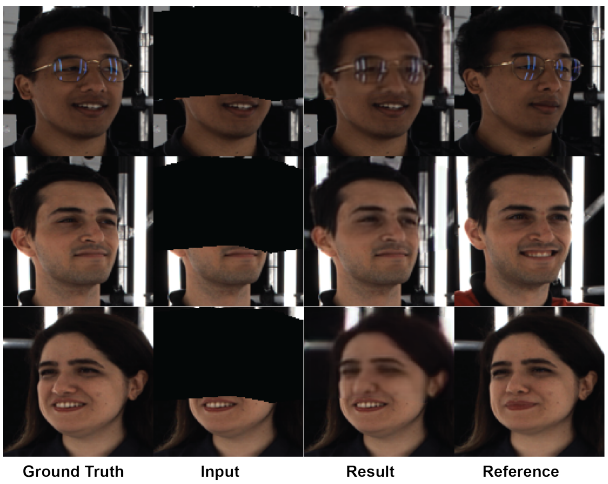

Fig. 9: Face completion of three different participants from HEADSET. The images are ground truth, input frame, inpainted result, and occlusion-free reference image, respectively.

Table 4: Accuracy of FEC models for our datasets (HEADSET-VoCap and HEADSET-LF) compared to available benchmark datasets for FEC.

Model	HEADSET-VoCap	HEADSET-LF	JAFFE [30]	AffectNet [32]	AFEW [11]	VGAF [50]
EfficientNet-B0, 8 classes [44]	58.19	61.98	46.00	60.10	55.14	68.29
EfficientNet-B2, 7 classes [46]	67.44	62.25	54.00	64.30	59.63	69.84
EfficientNet-B2, 8 classes [45]	62.46	61.50	54.00	60.90	57.78	70.23
ResMaskingNet, 7 classes [39]	51.11	53.34	46.95	49.81	–	–
Ad-Corre, 7 classes [13]	46.90	50.78	41.31	54.07	–	–

For preparing HMD masks on the video frames, we first created binary masks of the VR headset captured in experiment C. Then, we applied the masks to the ground truth images to be the input data for the inpainting network. We also used the first frames from each video sequence without any occlusion as a reference image that imposes an identity prior to the searching space of the network. The reference images were fed into the network jointly with the masked frames.

Samples of the qualitative results of the LGTSM model with a self-attention module on HEADSET are illustrated in Fig. 9. The examples are from three different individuals and captured by two distinct cameras. The first two rows are the inpainted results from starting frames, and the last row demonstrates an illustration of HMD removal outputs from the final frames. While the qualitative findings are promising, they exhibit temporal inconsistencies across frames, underscoring the need for additional research to comprehensively investigate potential strategies for mitigating HMD occlusion removal.

7 Conclusion

In this work, we have captured and presented a multimodal database that depicts humans performing posed and spontaneous facial expressions and subtle body movements. We have also recorded a part of the database with HMD occlusion. Our capturing setup includes a VoCap studio and a Lytro Illum camera. On top of the obtained textured meshes, colored point clouds, multi-view RGB-D images, and light field images, the raw captured data, calibration, and camera parameters are also made available. The proposed databases' performance in comparison with similar datasets has also been evaluated in different application scenarios. The provided material can facilitate the design of immersive media technologies and XR applications in which realistic human interaction is necessary. We believe that our database will then promote further research in data-driven techniques, computer vision for XR, human interactions in XR, and volumetric data reconstruction by providing a high-quality testing set for performance evaluation. Although the utilization of HEADSET holds significant potential for advancing research pertaining to XR applications, an extension of the existing HEADSET version can further enhance the progress of these technologies. This extension encompasses the inclusion of a greater number of participants, extended recording capabilities, as well as an examination of the impact of diverse factors such as age and medical conditions on individuals' perceived emotions.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956770. The data collection part was carried out with the support of Centre for Immersive Visual Technologies (CIVIT) research infrastructure, Tampere University, Finland. We want to especially thank Jani Käpylä, for his help during the capturing.

References

[1] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras. An integrated platform for live 3d human reconstruction and motion capturing. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):798–813, 2016.

[2] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8387–8397, 2018.

[3] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction*, pp. 279–283, 2016.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

[5] Y.-L. Chang, Z. Y. Liu, K.-Y. Lee, and W. Hsu. Learnable gated temporal shift module for deep video inpainting. *arXiv preprint arXiv:1907.01131*, 2019.

[6] A. Chatzitofis, L. Saroglou, P. Boutis, P. Drakoulis, N. Zioulis, S. Subramanyam, B. Kevelham, C. Charbonnier, P. Cesar, D. Zarpalas, et al. Human4d: A human-centric multimodal dataset for motions and immersive media. *IEEE Access*, 8:176241–176262, 2020.

[7] S.-Y. Chen, Y.-K. Lai, S. Xia, P. Rosin, and L. Gao. 3d face reconstruction and gaze tracking in the hmd for virtual interaction. *IEEE Transactions on Multimedia*, 2022.

[8] P. Chhokra, A. Chowdhury, G. Goswami, M. Vatsa, and R. Singh. Unconstrained kinect video face database. *Information Fusion*, 44:113–125, 2018.

[9] V. Chiesa and J.-L. Dugelay. On multi-view face recognition using lytro images. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2250–2254. IEEE, 2018.

[10] P. Cignoni, C. Rocchini, and R. Scopigno. METRO: Measuring error on simplified surfaces. *Computer Graphics Forum*, 17:167 – 174, 06 1998. doi: 10.1111/1467-8659.00236

[11] A. Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pp. 546–550, 2019.

[12] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016.

[13] A. P. Fard and M. H. Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022.

[14] C. Galdi, V. Chiesa, C. Busch, P. Lobato Correia, J.-L. Dugelay, and C. Guillemot. Light fields for face analysis. *Sensors*, 19(12):2687, 2019.

[15] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schönborn, and T. Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 75–82. IEEE, 2018.

[16] D. Girardeau-Montaut. Cloudcompare. *France: EDF R&D Telecom ParisTech*, 11, 2016.

[17] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pp. 117–124. Springer, 2013.

[18] D. Graziosi, O. Nakagami, S. Kuma, A. Zaghetto, T. Suzuki, and A. Tabatabai. An overview of ongoing point cloud compression standardization activities: video-based (v-pcc) and geometry-based (g-pcc). *APSIPA Transactions on Signal and Information Processing*, 9:e13, 2020. doi: 10.1017/ATSIP.2020.12

[19] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM*

Transactions on Graphics (ToG), 38(6):1–19, 2019.

[20] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017.

[21] HTC. Vive pro eye overview.

[22] A. S. Jackson, C. Manafas, and G. Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[23] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3334–3342, 2015.

[24] M. Kazhdan, M. Chuang, S. Rusinkiewicz, and H. Hoppe. Poisson surface reconstruction with envelope constraints. In *Computer Graphics Forum*, vol. 39, pp. 173–182. Wiley Online Library, 2020.

[25] M. Krivokuca, P. A. Chou, and P. Savill. 8i voxelized surface light field (8iVSLF) dataset. *ISO/IEC JTC1/SC29/WG11 MPEG, input document m42914*, 2018.

[26] C. Kyrlitsias and D. Michael-Grigoriou. Social interaction with agents and avatars in immersive virtual environments: A survey. *Frontiers in Virtual Reality*, 2:168, 2022.

[27] S. Li, W. Deng, and J. Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2852–2861, 2017.

[28] Y. Liu, Q. Yang, Y. Xu, and L. Yang. Point cloud quality assessment: Dataset construction and learning-based no-reference metric. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–26, 2023.

[29] J. Lou, Y. Wang, C. Nduka, M. Hamedi, I. Mavridou, F.-Y. Wang, and H. Yu. Realistic facial expression reconstruction for vr hmd users. *IEEE Transactions on Multimedia*, 22(3):730–743, 2019.

[30] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205. IEEE, 1998.

[31] R. Min, N. Kose, and J.-L. Dugelay. Kinectfacedb: A kinect database for face recognition. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 44(11):1534–1548, 2014.

[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[33] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 343–352, 2015.

[34] N. Numan, F. Ter Haar, and P. Cesar. Generative rgb-d face completion for head-mounted display removal. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 109–116. IEEE, 2021.

[35] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 741–754, 2016.

[36] R. Pagés, K. Amplianitis, J. Ondrej, E. Zerman, and A. Smolic. Volograms & v-sense volumetric video dataset. *ISO/IEC JTC1/SC29/WG07 MPEG2021/m56767*, 2021.

[37] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

[38] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9054–9063, 2021.

[39] L. Pham, T. H. Vu, and T. A. Tran. Facial expression recognition using residual masking network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 4513–4519. IEEE, 2021.

[40] R. Raghavendra, K. B. Raja, and C. Busch. Exploring the usefulness of light field cameras for biometrics: An empirical study on face and iris recognition. *IEEE Transactions on Information Forensics and Security*, 11(5):922–936, 2015.

[41] I. Reimat, E. Alexiou, J. Jansen, I. Viola, S. Subramanyam, and P. Cesar. Cwipc-sxr: Point cloud dynamic human dataset for social xr. In *Proceedings of the 12th ACM Multimedia Systems Conference*, pp. 300–306, 2021.

[42] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[43] S. Saito, T. Simon, J. Saragih, and H. Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 84–93, 2020.

[44] A. V. Savchenko. Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In *2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 119–124. IEEE, 2021.

[45] A. V. Savchenko. Video-based frame-level facial analysis of affective behavior on mobile devices using efficientnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2359–2366, 2022.

[46] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing*, 2022.

[47] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. Cohen, M. Krivokuća, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. Tourapis, and V. Zakharchenko. Emerging MPEG Standards for Point Cloud Compression. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(1):133–148, 2019. doi: 10.1109/JETCAS.2018.2885981

[48] A. Sepas-Moghaddam, V. Chiesa, P. L. Correia, F. Pereira, and J.-L. Dugelay. The ist-eurecom light field face database. In *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6. IEEE, 2017.

[49] A. Sepas-Moghaddam, A. Etemad, F. Pereira, and P. L. Correia. Capsfield: light field-based face and expression recognition in the wild using capsule routing. *IEEE Transactions on Image Processing*, 30:2627–2642, 2021.

[50] G. Sharma, A. Dhall, and J. Cai. Audio-visual automatic group affect analysis. *IEEE Transactions on Affective Computing*, 2021.

[51] Z. Su, L. Xu, Z. Zheng, T. Yu, Y. Liu, and L. Fang. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *European Conference on Computer Vision*, pp. 246–264. Springer, 2020.

[52] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro. Geometric distortion metrics for point cloud compression. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3460–3464, 2017. doi: 10.1109/ICIP.2017.8296925

[53] I. A. Verpaalen, G. Bijsterbosch, L. Mobach, G. Bijlstra, M. Rinck, and A. M. Klein. Validating the radboud faces database from a child's perspective. *Cognition and Emotion*, 33(8):1531–1547, 2019.

[54] M. Wang, X. Wen, and S.-M. Hu. Faithful face image completion for hmd occlusion removal. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 251–256. IEEE, 2019.

[55] Z. Yu, J. S. Yoon, I. K. Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2990–3000, 2020.

[56] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019.

[57] J. Zhang, D. Huang, Y. Wang, and J. Sun. Lock3dface: A large-scale database of low-cost kinect 3d faces. In *2016 International Conference on Biometrics (ICB)*, pp. 1–8. IEEE, 2016.

[58] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[59] Y. Zhao, W. Chen, J. Xing, X. Li, Z. Bessinger, F. Liu, W. Zuo, and R. Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018.

[60] H. Zheng, W. Wang, F. Wen, and P. Liu. A complementary fusion strategy for rgb-d face recognition. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*, pp. 339–351. Springer, 2022.