IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 18, 2020, accepted September 21, 2020, date of publication September 23, 2020,
date of current version October 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3026276

HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media

ANARGYROS CHATZITOFIS[1,2], (Graduate Student Member, IEEE), LEONIDAS SAROGLOU[2], PRODROMOS BOUTIS[2], PETROS DRAKOULIS[2], NIKOLAOS ZIOULIS[2], SHISHIR SUBRAMANYAM[3], (Graduate Student Member, IEEE), BART KEVELHAM[4], CAECILIA CHARBONNIER[4], PABLO CESAR[3], (Senior Member, IEEE), DIMITRIOS ZARPALAS[2], STEFANOS KOLLIAS[1], (Fellow, IEEE), AND PETROS DARAS[2], (Senior Member, IEEE)

[1]School of Electrical and Computer Engineering, National Technical University of Athens, 157 73 Athens, Greece
[2]Information Technologies Institute, Centre for Research & Technology Hellas, 57001 Thessaloniki, Greece
[3]Centrum Wiskunde & Informatica, 1098 Amsterdam, The Netherlands
[4]Artanim Foundation, 1217 Geneva, Switzerland

Corresponding author: Anargyros Chatzitofis (tofis@iti.gr)

This work was supported by the European Union (EU) funded project VRTogether H2020 under Agreement 762111.

ABSTRACT We introduce HUMAN4D, a large and multimodal 4D dataset that contains a variety of human activities simultaneously captured by a professional marker-based MoCap, a volumetric capture and an audio recording system. By capturing 2 female and 2 male professional actors performing various full-body movements and expressions, HUMAN4D provides a diverse set of motions and poses encountered as part of single- and multi-person daily, physical and social activities (jumping, dancing, etc.), along with multi-RGBD (mRGBD), volumetric and audio data. Despite the existence of multi-view color datasets captured with the use of hardware (HW) synchronization, to the best of our knowledge, HUMAN4D is the first and only public resource that provides volumetric depth maps with high synchronization precision due to the use of intra- and inter-sensor HW-SYNC. Moreover, a spatio-temporally aligned scanned and rigged 3D character complements HUMAN4D to enable joint research on time-varying and high-quality dynamic meshes. We provide evaluation baselines by benchmarking HUMAN4D with state-of-the-art human pose estimation and 3D compression methods. We apply OpenPose and AlphaPose reaching 70.02% and 82.95% $mAP_{PCKh\text{-}0.5}$ on single- and 68.48% and 73.94% $mAP_{PCKh\text{-}0.5}$ on two-person 2D pose estimation, respectively. In 3D pose, a recent multi-view approach named Learnable Triangulation, achieves 80.26% $mAP_{PCK3D\text{-}10cm}$. For 3D compression, we benchmark Draco, Corto and CWIPC open-source 3D codecs, respecting online encoding and steady bit-rates between 7-155 and 2-90 Mbps for mesh- and point-based volumetric video, respectively. Qualitative and quantitative visual comparison between mesh-based volumetric data reconstructed in different qualities and captured RGB, showcases the available options with respect to 4D representations. HUMAN4D is introduced to enable joint research on spatio-temporally aligned pose, volumetric, mRGBD and audio data cues. The dataset and its code are available online.

INDEX TERMS Dataset, 4D, multi-view, motion capture, RGBD, volumetric video, pose estimation, 3D compression, 4D capture, visual evaluation, benchmarking, depth sensing, audio, social activities.

I. INTRODUCTION

Inhabitance in a 4D world of moving 3D objects of various shapes and colors increases the need to capture and extensively study, analyze and exploit the 4D data around us, especially now, with the massive development of low-cost sensing devices [1]. Nowadays, volumetric video of humans, captured with the aid of multiple cameras, and scanned 3D characters, animated with the use of motion capture (MoCap) technologies, comprise the core elements for human-centric 4D media production, a domain essential in several technological and industrial sectors.

On the one hand, these technologies constitute key elements in immersive experiences that provide remote

The associate editor coordinating the review of this manuscript and approving it for publication was Nilanjan Dey.

virtual presence and co-presence (e.g. XR conferencing [2], XR museums [3], etc.). The experiences are further enhanced by augmenting the virtual and immersive worlds with photo-realistic representations that enable highly natural and realistic audiovisual communication between multiple users.

On the other hand, dense 4D data cues produced with such technologies contain space-time coherent information of shape, motion, and appearance of people, attracting the interest of the computer vision research community and beyond. Several research works [4], [5] provide large corpora with synthetic humans generated based on human body priors [6], motion capture data and more. By applying 3D surface reconstruction methods [7]–[14] on 3D or 4D data captured with single or multiple spatio-temporally aligned RGBD sensors, volumetric video is reconstructed in either real-time or offline. Fusing volumetric video with high quality 3D scans and motion capture enables the study and development of data-driven approaches across several domains, such as 2D human pose estimation [15]–[19], 3D pose estimation [20]–[26], motion analysis [27], [28], 3D/4D volumetric reconstruction [7]–[13], [29], performance capture [30], [31], volumetric video compression [32]–[36], photorealistic representations [14] and more.

The advancement of shape and motion computer vision techniques, the development of immersive media technologies, as well as the interest of the industry in human-centric 4D media production, highly and rapidly increase the need for large, high-quality datasets that will act as cornerstones for their continuous development, also enabling their joint evolution. Nevertheless, at the moment, only few datasets are partially focused on some of the aspects of these challenging tasks.

On top of that, several computer vision methods approach 3D/4D research tasks from monocular or HW-SYNCed multi-view color (i.e. 2D) streams. However, by definition, 2D data cannot cope with the intricacies of 3D/4D shape or form, at least to the extent that the volumetric data can. That is probably due to the lack of HW-SYNCed depth/volumetric data from public resources. For instance, the lack of HW-SYNCed volumetric data along with ground-truth 3D poses for supervision eliminates the attempts for data-driven 3D pose estimation approaches from volumetric data.

To this end, we create HUMAN4D, a dataset that fills these gaps by providing professional motion capture along with volumetric data captured in 3D character and mesh- and point-based volumetric representations. In particular:

- We introduce a publicly available 4D dataset containing a large corpus of annotated spatio-temporally aligned multi-view RGBD (mRGBD), volumetric and motion capture data, in order to enable extensive research on several computer vision and graphics topics.
- To the best of our knowledge, HUMAN4D is the first dataset that provides HW-SYNCed mRGBD frames along with marker-based motion capture and audio data cues, with the use of recent consumer-grade depth sensing devices, cutting-edge optical motion capture technologies and body-worn audio recording, respectively.
- We provide pose estimation baselines by applying data-driven 2D and 3D pose estimation algorithms on single- and multi-view data sequences, along with insights with respect to the advantages of HUMAN4D for training such methods.
- We perform and report a detailed study on volumetric data compression using 3D codecs, examining the rate distortion from several perspectives, while respecting online volumetric video encoding and steady bit-rates.
- We conduct and report objective visual quality evaluation on various volumetric representations, i.e. mesh-based volumetric data evaluation across various reconstruction qualities.

The remainder of this paper is organized as follows: Sec. II overviews related datasets including 4D data in a similar aspect; Sec. III describes in detail the HUMAN4D dataset, giving evidence with respect to its creation and statistics; Sec. IV benchmarks 2D and 3D pose estimation data-driven models on HUMAN4D; while Sec. V benchmarks 3D codecs and compares mesh-based 4D representations with respect to visual quality using well-known objective metrics; in Sec. VI, we discuss the impact of this dataset to the research community and beyond; finally, Sec. VII concludes the paper and discusses future work.

II. RELATED WORK

Over the past few decades, the computer vision research community has showed an increased interest for virtual human related technologies. A variety of traditional and learning-based computer vision methods are targeting open research problems using motion, volumetric, image and action-based data. In this section, we discuss relevant datasets [37]–[42], providing details and explaining the nature of the data they offer to the research community. A brief overview of these datasets follows, while Table 1 summarizes their features and modalities.

MHAD [37]: One of the first publicly available datasets offering MoCap and RGBD data is (Berkeley) MHAD. The MHAD dataset contains spatio-temporally aligned data cues captured with a professional MoCap system with active markers [43] along with 12 RGB and 2 MS Kinect v2 (RGBD) cameras, 6 wearable inertial sensors (accelerometers only) and 4 microphones, recording the audio signals during the performance of the actions. The dataset consists of 659 data sequences from 11 human actions performed by 12 subjects. Although MHAD enables research on multi-view pose estimation and beyond, the MS Kinect v2 devices are only 2 and not HW-SYNCed, resulting in the existence of spatio-temporal offsets between the deprojected depth maps (point-clouds) and the 3D poses of the MoCap, limiting that way the joint use of 3D pose and volumetric data.

Human3.6M [38]: Human3.6M (H36M) contains a huge corpus with 3.6 million 3D human poses of 5 female and

TABLE 1. Summary of state-of-the-art datasets and HUMAN4D with respect to the available features and modalities.

	MHAD$_{(2013)}$ [37]	Human3.6M$_{(2014)}$ [38]	CMUPanoptic$_{(2015)}$ [39]	HUMBI$_{(2018)}$ [40]	HUMAN4D$_{(2020)}$
Body Pose	✓	✓	✓	✓	✓
Marker-based MoCap	✓	✓	✗	✗	✓
Body Part Segments	✗	✗	✗	✓	✗
Multi-view RGB	✓	✓	✓	✓	✓
Multi-view Depth	✓	✗	✓	✗	✓
3D Meshes	✗	✗	✗	✓	✓
Point-clouds	✗	✗	✓	✗	✓
Audio Cues	✓	✗	✗	✗	✓
Gaze Features	✗	✗	✗	✓	✗
Hand Features	✗	✗	✓	✓	✗
Facial Features	✗	✗	✓	✓	✗
Rigged Characters	✗	✓	✗	✗	✓
Multi-person	✗	✗	✓	✗	✓

6 male subjects. Similarly to HUMAN4D, the subjects perform a set of motions and poses (captured with 10 motion capture cameras) from daily human activities (taking photos, talking on the phone, eating, sitting, etc.), along with synchronized color images from 4 synchronized color cameras, depth maps from 1 single Time-of-Flight (ToF) depth sensor and accurate 3D body scans of the subject actors involved. H36M constitutes one of the most widely used datasets for human-centric computer vision research tasks, however, there still exist some drawbacks. Only the color cameras support hardware inter-synchronization, there is only one depth sensor with low depth map resolution, while the set of motion capture cameras is limited (10) in comparison with HUMAN4D (24). Finally, the recent human-centric research advances and efforts are focused on multi-person captures (e.g. including social activities) similar to ones provided by HUMAN4D and other datasets [39], [40], contrary to H36M which contains only single-person sequences.

CMUPanoptic [39]: CMUPanoptic (CMU) is the largest public dataset in terms of the number of camera views (521), capturing natural interactions of up to 8 subjects performing social activities with uncontrolled behaviour and appearance. The dataset has been captured using the Panoptic Studio [39], a massively multi-view capture system consisting of 480 VGA, 31 HD and 10 RGBD (Kinect v2) cameras, distributed over the surface of a geodesic sphere. Beyond body poses, CMU also contains 3D facial landmarks and 2D/3D hand pose data cues. Even though CMU currently constitutes one of the richest publicly available datasets in the field, HUMAN4D enables further research perspectives. Despite its spatio-temporal setting, CMU does not provide HW volumetric synchronization since the time alignment between the Kinect v2 RGBD streams is achieved through a hardware modification using the microphone array of each device, incapable to provide synchronization precision comparable to HUMAN4D (see Sec. III-B1). Finally, the pose estimates have not been captured using a professional marker-based motion capture solution as in HUMAN4D; instead, an accurate marker-less approach has been used.

HUMBI [40]: Another large and publicly available multi-view dataset is HUMBI, focusing on human body expressions with natural clothing, aiming to facilitate modeling of view-specific appearance and geometry of gaze, face, hand, body, and garment from several and various people. HUMBI complements the publicly available datasets with respect to the number of camera views (107 synchronized HD cameras) and subjects (772 distinctive subjects across gender, ethnicity, age, and physical condition). The dataset includes five elementary body expressions, i.e. gaze, face, hand, body and garment. With the use of SMPL [6], HUMBI provides mesh-based 3D geometry of the subjects along with their respective texture atlases. For HUMBI, the use of depth sensors was out of scope, thus multi-view depth sensing was not considered.

HUMAN4D aims to tackle lacking areas of existing, publicly available 4D datasets. HUMAN4D consists of a large corpus of spatio-temporally aligned mRGBD, volumetric and motion capture data cues, providing high synchronization precision between the multiple RGBD streams exploiting the HW-SYNC capabilities of the sensors. On top of that, HUMAN4D contains (social) activities between multiple subjects (2), enabling research on challenging computer vision tasks under the multi-person aspect (e.g. occlusions, multiple person instances in the field of view, larger volumetric areas, etc.). HUMAN4D is meant to provide the computer vision research community with data that will enable the research and development of novel approaches on intensively active human-centric research domains. It is worth noting that the consumer-grade depth sensing devices used for the RGBD data capturing are commercially available in the market, allowing the experimentation and development of computer vision algorithms applicable even for production purposes.

III. HUMAN4D DATASET

A. 4D CAPTURING SETTING

The capturing of the dataset took place in a professional motion capture studio (Artanim Foundation[1]) where, beyond the motion capture system, special portable equipment for volumetric capturing was set up, as depicted in Fig. 1. In particular, 24 motion capture (MoCap) cameras along with 4 stereo-based depth sensors and microphones using HW and software (SW) synchronization (see Sec. III-C1 for details) were used, to capture the whole dataset. All 24 motion capture

[1] http://artanim.ch/

FIGURE 1. Pictures taken during the preparation and capturing of the HUMAN4D dataset (in Artanim's facilities). The room is equipped with 24 Vicon MXT40S cameras rigidly placed on the walls, a portable volumetric capturing system with 4 Intel RealSense D415 depth sensors temporarily set up to capture the RGBD data cues and wearable microphones for the actors.

cameras were rigged on the walls, to maximize the effective experimentation volume. The high number of motion cameras (24) increases the accuracy of the motion capture due to the elimination of occlusions, providing that way high precision ground-truth poses for the dataset. The actual capturing space was set in an area of approximately $4m \times 4m$ so that the bodies of the actors were at least partially in the field-of-view of the RGBD cameras during the performances. These cameras were placed at the 4 corners of the stage in a cross schema. The floor-plan of the whole capturing setup is illustrated in Fig. 2. Finally, a 3D body scanner was used to obtain an accurate 3D mesh-based volumetric model of one of the actors.

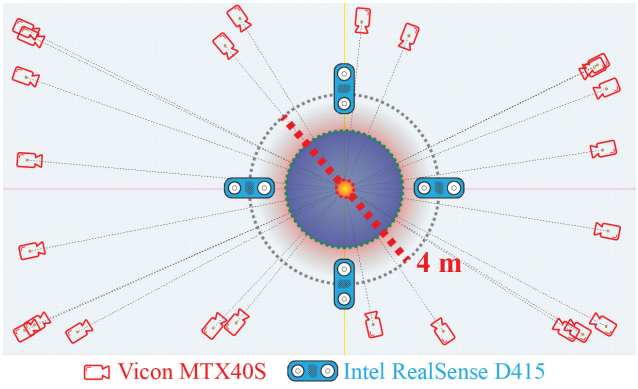

FIGURE 2. Capturing space floor-plan showing the poses of 24 Vicon MXT40S cameras and 4 Intel RealSense D415 sensors.

B. DATASET CREATION

For the creation of the dataset, 4 professional actors, 2 female and 2 male were recruited, in order to pursue the highest possible quality of the captured actions, with respect to the authenticity of the performances. Within HUMAN4D, without the post-processing products (i.e. volumetric data), we captured and introduce the following:

- Multimodal data of 14 single-person and 5 two-person actions (19 in total), including physical exercises, daily and social activities, totalling 56 single-person and 10 two-person sequences, respectively. In Table 2, details with respect to HUMAN4D activities are figured.
- Projection matrices and external calibration camera parameters retrieved using an anchor-based calibration method to reduce pairwise accumulating errors, enabling 2D projection of 4D data to the various camera views and vice versa.
- 30 audio cues for some of the activities where the actors had to talk and act based on specific scripts and scenarios (see Table 2).

TABLE 2. Details with respect to HUMAN4D physical, daily and social activities.

	activity	# frames	audio	type
Single-person	*running*	2,050	✗	physical
	jumping_jack	1,974	✗	physical
	bending	2,156	✗	physical
	punching_n_kicking	2,079	✗	physical
	basketball_dribbling	2,124	✗	physical
	laying_down	4,082	✗	physical
	sitting_down	3,288	✗	daily
	sitting_on_a_chair	2,797	✗	daily
	talking	2,377	✓	daily
	object_dropping_n_picking	1,768	✗	daily
	stretching_n_talking	2,787	✓	physical
	talking_n_walking	2,889	✓	daily
	watching_scary_movie	2,194	✓	daily
	in-flight_safety_announcement	6,192	✓	daily
Multi-person	*watching_football_together*	1,760	✓	social
	dancing_together	1,356	✓	social
	physical_examination	2,328	✓	social
	whispering	3,045	✓	social
	card_trick	3,060	✓	social
		50,306		

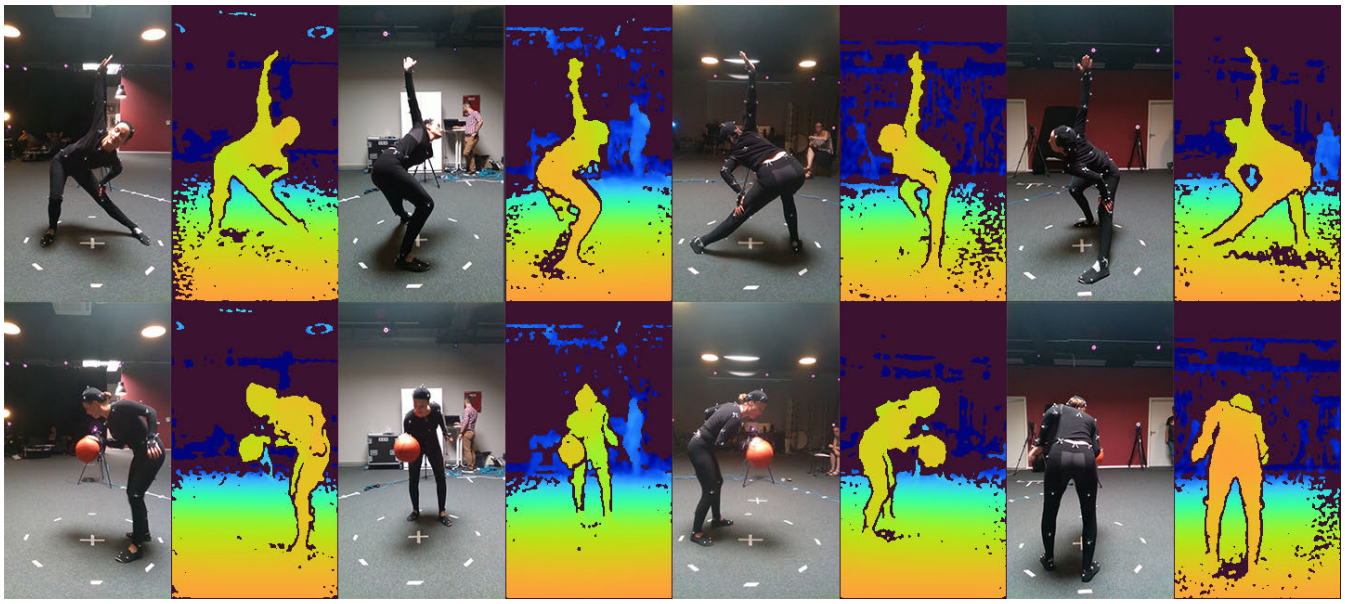

FIGURE 3. HW-SYNCed multi-view RGBD samples (4 RGBD frames each) from *"stretching_n_talking"* (top) and *"basketball_dribbling"* (bottom) activities. The depth maps are colorized using TURBO colormap [44].

- Synchronization between the modalities by providing timestamped data.
- 1 scanned and rigged 3D model of one of the professional actors.
- A set of benchmarks to facilitate comprehensive evaluation of 2D and 3D pose estimation methods, along with evaluation of volumetric video production and compression quality.

Following, we describe in detail the modalities we used and the techniques we applied to capture and create the dataset.

1) SPATIO-TEMPORALLY ALIGNED mRGBD CAPTURE

To the best of our knowledge, HUMAN4D is the first publicly available dataset that offers HW synchronized multi-view RGBD data captured in a real-time manner. Most of the existing datasets use synchronized RGB cameras [38] or previous versions of Microsoft Kinect for RGBD capturing [39], which do not support HW triggering, requiring SW-based soft synchronization solutions.

In HUMAN4D, we instead use the Intel RealSense D415 sensor which offers this functionality [45]. D415 sensors can be configured in either master or slave synchronization mode, eliminating the need for external HW triggering when connected in a device cluster. One device can be set as ''master'', providing the synchronization signal, and the rest as ''slaves'' that receive it and cohere. The impact of HW-SYNCed mRGBD capture for volumetric- and pose-related tasks is depicted in Fig. 6, where point-clouds extracted by deprojecting mRGBD frames from HUMAN4D and CMU [39] are compared, showcasing the improved temporal alignment of the HW-SYNCed HUMAN4D against CMU data. It is worth noting that CMU constitutes currently the only existing dataset that provides synchronized depth maps by applying a HW modification on the Kinect v2 devices.

Regarding depth capturing, the sensors were used in ''high accuracy'' mode, offering only the high confidence depth estimates, therefore producing accurate but sparse depth data. It is worth noting that we configured the sensors exploiting their spatial filtering and exposure adjustment capabilities to capture the best possible depth quality. We captured the mRGBD data using the capturing system[2] proposed by Sterzentsenko *et al.* [46], while spatial alignment between the sensors was achieved using the multi-sensor calibration schema proposed by Papachristou *et al.* [47]. HW-SYNCed mRGBD samples are depicted in Fig. 3.

2) 3D SCANNED AND RIGGED CHARACTER

To obtain an animatable mesh, one of the actors was scanned using a custom photogrammetry-based body scanning rig (Fig. 4). The rig consisted of 96 Canon Powershot A1400 cameras controlled using SW-based on the Canon Hack Development Kit (CHDK) [49]. Lighting was provided by LED strips mounted on the rig. All cameras were triggered in a synchronized manner. To aid the photogrammetric reconstruction of the bodyscan, the dark MoCap suit worn by the actor was temporarily augmented with colored paper markers, which were removed before the MoCap process.

Using a commercial photogrammetry SW tool, Agisoft Metashape [48], the individual photos were aligned to reconstruct a textured 3D mesh. After the cleanup of mesh artifacts from the reconstruction process, the mesh was rigged and skinned for animation, using a standard full-body humanoid skeleton created by a professional 3D animator.

[2]https://github.com/VCL3D/VolumetricCapture

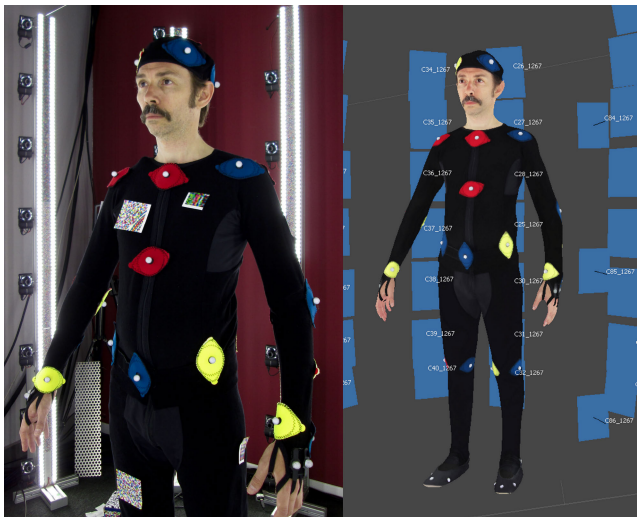

FIGURE 4. Using a custom photogrammetry rig with 96 cameras, photos were taken of the actor (left) and reconstructed into a 3D textured mesh using Agisoft Metashape [48] (right).

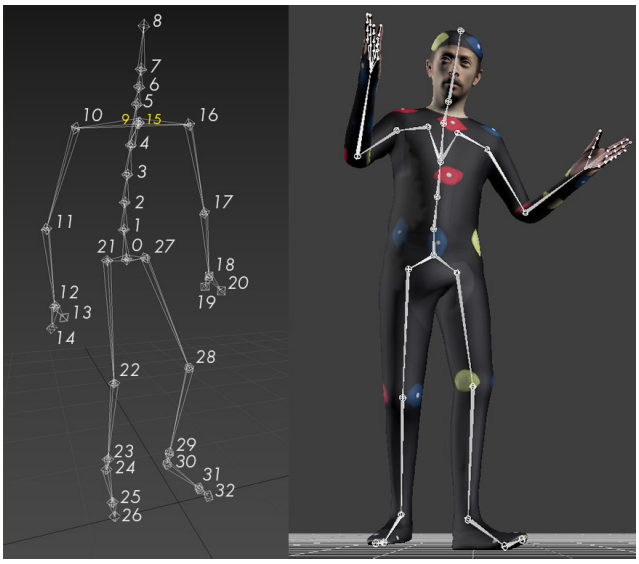

FIGURE 5. (Left) Initial MoCap skeleton structure mapped to 3D and 2D pose joint indices. (Right) Animations for the scanned actor were re-targeted to match the skeleton rig of the 3D model.

3) OPTICAL MARKER-BASED MOTION CAPTURE

To obtain reference animation of the 4 actors performing the various activities, a professional motion capture setup was used. The setup consisted of 24 Vicon MXT40S cameras (Vicon, Oxford Metrics, UK) sampling at 120Hz. Each actor wore a dedicated motion capture suit with 53 attached retro-reflective markers. This dense marker set along with the high number of motion cameras (24) allowed us to capture highly accurate and precise MoCap data to serve as ground-truth for training, supervising and evaluating data-driven approaches and beyond.

For the purpose of subject calibration, each actor was asked to perform a full range of motion of all joints. The procedure ensured that the joint locations were correctly mapped to the set of the tracked markers. Before each activity, the actors were asked to start in a T-pose and then proceed to their assigned activity.

The captured animations of the actor whose body was subsequently scanned, underwent a retargeting process by a professional 3D animator. The goal of this process was to adjust the recorded animations to where slight differences between the captured MoCap skeleton structure and the one of the rigged 3D model exist, as illustrated in Fig. 5 (Right).

4) AUDIO RECORDING

The use of audio and its fusion with visual data have shown significant results in various research tasks such as human emotion recognition [50], scene analysis [51], human activity recognition [52] and more. To this end, also targeting the capture of social activities, we recorded audio during the performance of some of the actions. In particular, 30 of the activities (see Table 2) include audio either as a monologue (single-person) or conversation between two subjects, based on the related scripts and scenarios. For this purpose, wireless body-worn microphones were used to record the audio cues. The audio recording was performed at the frequency of 48 kHz.

C. DATASET PROCESSING AND ANNOTATIONS

1) SYNCHRONIZATION AND CALIBRATION

Inter- and intra-modality synchronization is a prerequisite for such datasets. The motion capture cameras operate in inter-camera synchronization by default. With respect to the mRGBD capturing setting, as we already mentioned, Intel RealSense D415 sensors offer intra- and inter-sensor HW synchronization as well. With respect to the inter-modality synchronization, considering the motion capture clock as reference for the full system, along with the mRGBD and audio data timestamping, a SW-based synchronization technique was applied to temporally align the data. In particular, given the motion capture frequency equal to 120 Hz, the temporally closest MoCap sample to every mRGBD frame timestamp was considered the matching pose, giving a low temporal difference t_d, where $t_d \leq \frac{1}{120}/2$ ms $\implies$ $t_d \leq 4.16$ ms. The initial temporal offset between the modalities was detected with the use of a marker-equipped (2 markers) clapperboard at the beginning of each sequence, enabling all the modalities to capture the time instance of the clapping event. In detail, for the motion capture data sequences, the 3D position signals of the clapperboard markers were analyzed to detect the clap event by identifying the time instance when the euclidean distance between the markers is the minimum; for the audio signals, the clap event caused an easily detectable peak on the amplitude of the audio signals, while for the RGBD data, the event was manually detected.

For the spatial alignment of the modalities, the MoCap system was calibrated once before the captures, while the

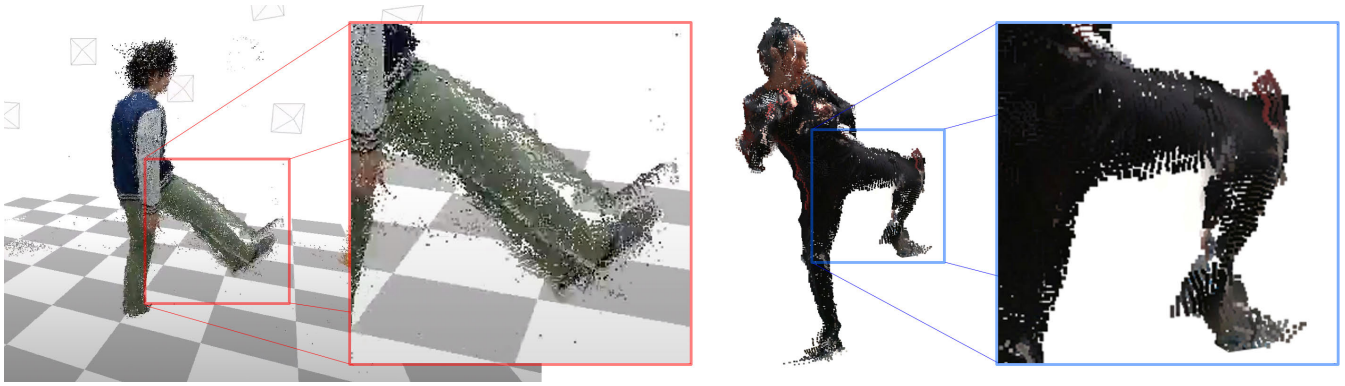

FIGURE 6. Colored point-clouds from CMU [39] (Left) and HUMAN4D (Right) datasets showcase the benefits of HW-SYNC. In CMU, where the Kinect devices are modified for synchronization purposes, the leg of the subject is corrupted in a slow movement (i.e. slow leg lifting) due to the existence of temporal offsets between the devices. In HUMAN4D, the leg is appropriately captured in a fast movement (i.e. punching and kicking).

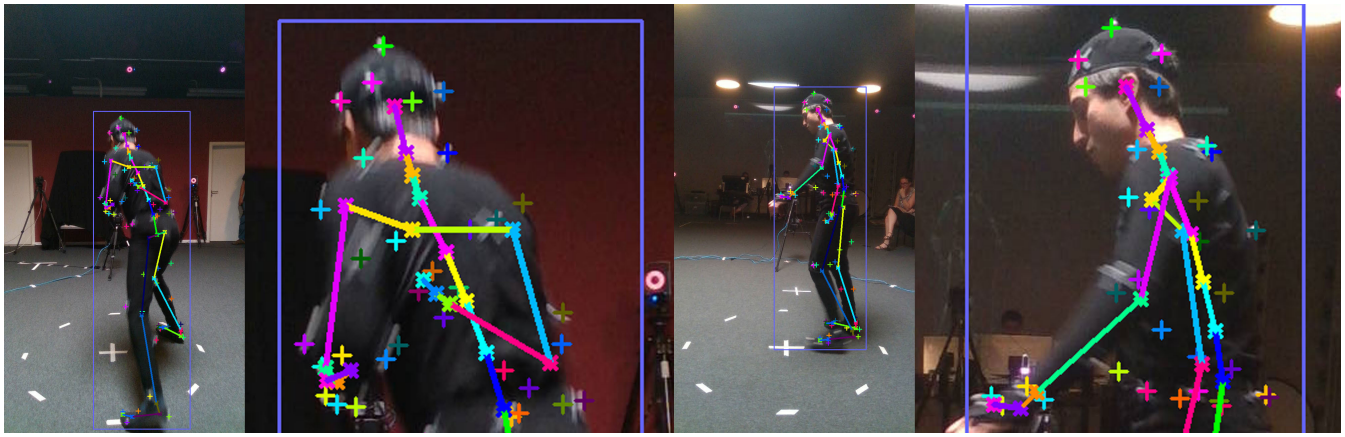

FIGURE 7. Joint and marker 3D positions projected on color views with high accuracy. The marker projection accuracy which is clearly visible on the color views showcases the precision of the spatio-temporal alignment between the 3D poses (MoCap) and the RGBD data.

mRGBD system was calibrated per subject (every subject performed all the actions at once). The spatial alignment between MoCap and mRGBD was achieved by applying a semi-automatic technique, capturing short sequences of moving retro-reflective markers using both modalities before the capturing of each subject. For these sequences, the infrared (IR) stream of the sensors was enabled instead of the color. The details of the inter-modality spatial calibration go beyond the scope of this paper.

2) 2D AND 3D POSE FROM MOTION CAPTURE

The spatio-temporal alignment between the modalities and the highly frequent and precise 3D motion capture enable the extraction of 3D poses accurately mapped on the RGBD data cues. With a set of $J = 33$ j-joints, as depicted in Fig. 5, a 3D pose per frame f and skeleton s is mapped to every single mRGBD frame. Then, by applying inverse transformation per camera pose and projecting the 3D positions of the joints on the RGBD views, the 2D keypoints $\mathcal{K}$ are calculated by:

$$\mathcal{K}(f, s, j) = \pi(\mathbf{T}_{g\to l}(x_{f,s,j}), \mathbf{K}_s), \tag{1}$$

where $x_{f,s,j} \in \mathbb{R}^3$ is the 3D position of joint j, $\mathbf{T}_{g\to l}$ is the transformation from the global (g) coordinate system to the local (l) one of sensor s with the arrow showing the direction of the transformation. π denotes the projection function that transforms the 3D coordinates to pixels, using sensor's intrinsic parameters matrix $\mathbf{K_s}$. The 2D outcomes of this processing are depicted in Fig. 7 and 8.

Furthermore, considering the MoCap marker 3D positions and their corresponding 2D projections on the sensor views (using the projection of Eq. (1)), we extract the 3D and 2D bounding boxes containing each subject per frame, by fitting a rectangular slightly padded (2% of the dimension size per side) prism and box around the 3D positions and 2D projections, respectively.

3) VOLUMETRIC DATA FROM MULTI-VIEW RGBD

Real-time 4D reconstruction evolves as a cutting-edge component in XR applications and beyond, especially focused on challenging dynamic data such as rigid and non-rigid human motions. Key concept of this dataset is the exploitation of the mRGBD cues of human activities to produce and dispose

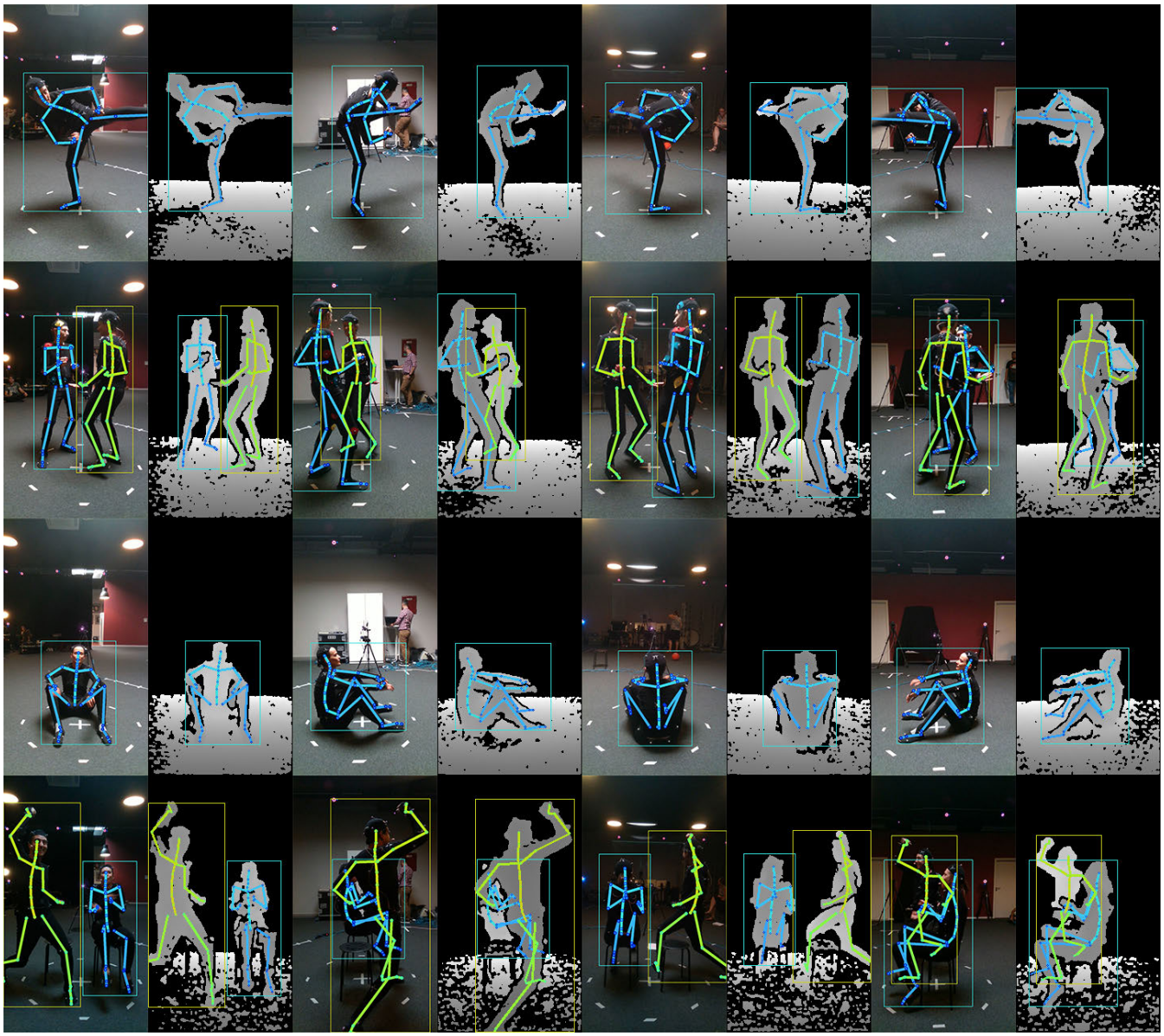

FIGURE 8. 2D poses and bounding box annotations illustrated on various color and depth frames. The rows depict the 4 different views of mRGBD frames both from single- and two-person activities.

volumetric data captured in a real-time manner, in the form of colored point-cloud and colored/textured 3D mesh instances for every single mRGBD frame.

Point-Cloud: An RGBD image is composed of a color image $\mathcal{I}$ and a depth image $\mathcal{D}$, which, after the application of a local transformation between them, are registered to the same coordinate frame. Then, given the depth sensors poses ($\mathbf{T}_v := \begin{bmatrix} \mathbf{R}_v & \mathbf{t}_v \\ \mathbf{0} & 1 \end{bmatrix}$) known in a common coordinate system, where $\mathbf{R}_s$ and $\mathbf{t}_s$ denote rotation and translation, respectively, we transform every depth pixel p, $p \in \mathcal{D}_s$, from the depth image domain coordinates of each view to a global coordinate system by:

$$\mathcal{T}_{l\to g}(p) = \mathbf{T}_{l\to g}\pi^{-1}(D_s(p), \mathbf{K}_s, p), \tag{2}$$

where $\mathbf{T}_{l\to g}$ is the relative pose from the local (l) coordinate system of sensor s to the global (g) one with the arrow showing the direction of the transformation. π^{-1} denotes the deprojection function that transforms the pixel to 3D coordinates, using sensor's intrinsic parameters matrix $\mathbf{K_s}$. Merging the transformed partial point clouds from each view to the global space, results in the colored point cloud data. The outcome of this process is illustrated in Fig. 9.

3D Mesh: Beyond point-based volumetric data, watertight colored and textured 3D mesh instances are reconstructed in a real-time manner (up to the frequency of the sensor acquisition, i.e. 30 fps) applying the GPU-based implementation proposed by Alexiadis *et al.* [8], based on the fast Fourier Transform (FFT) -based approach proposed by Kazhdan [53].

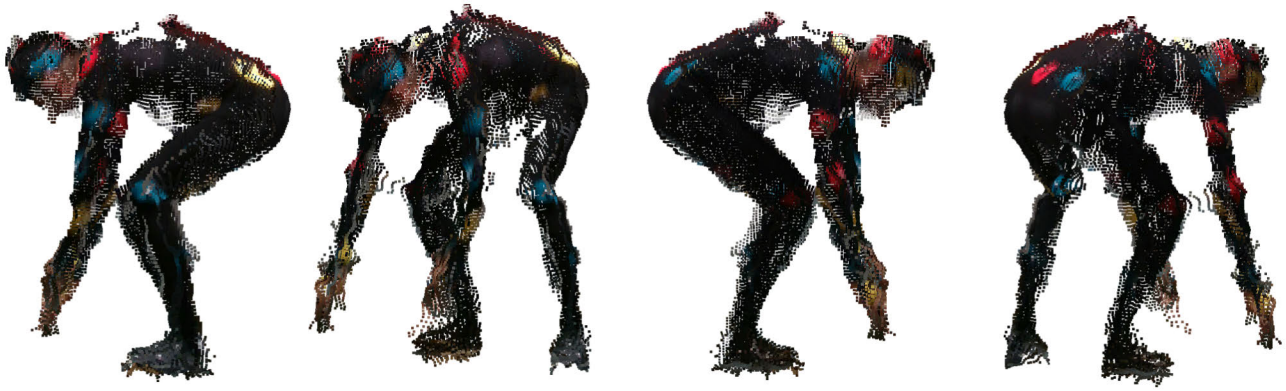

FIGURE 9. Merged reconstructed point-cloud from one single mRGBD frame from various views.

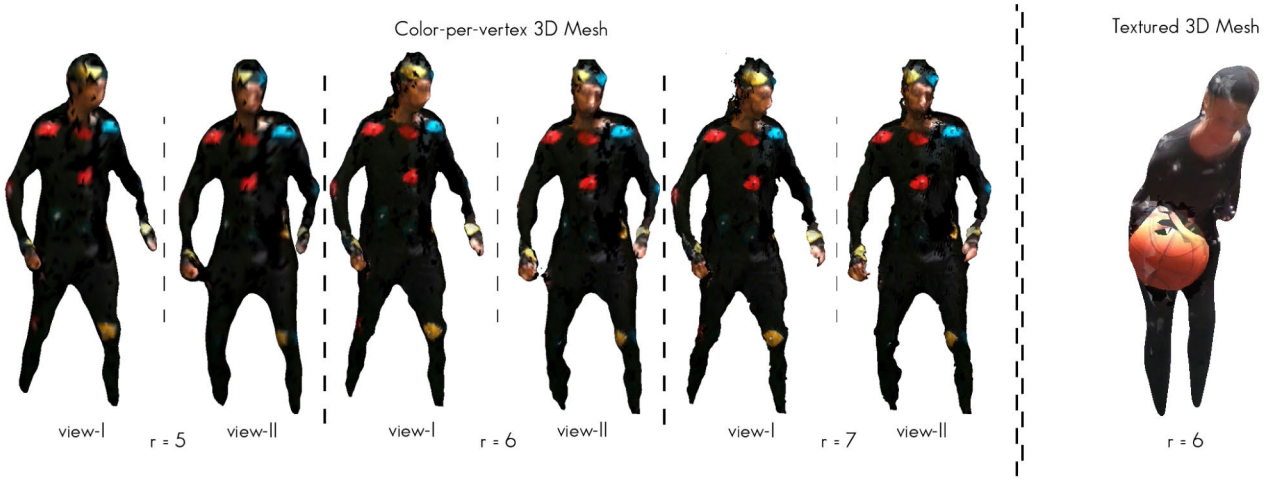

FIGURE 10. Reconstructed [8] mesh-based volumetric data with (Left) color per vertex visualization in 3 voxel-grid resolutions, i.e. $r = 5$, $r = 6$ and $r = 7$ and (Right) textured 3D mesh sample in voxel-grid resolution for $r = 6$.

The 3D geometry reconstruction relies on a scalar volume function $V(q)$ containing the splatted 3D surface information, as given by the point cloud calculated using the depth maps, defined over a 3D grid $q = [q_X; q_Y; q_Z]_T \in \{1, \ldots, N_X\} \times \{1, \ldots, N_Y\} \times \{1, \ldots, N_Z\}$, inside the foreground object's bounding box. This 3D grid of $V(q)$ is considered the volume resolution of the 3D reconstruction, used with power of 2 components for FFT, i.e. $2^r \times 2^{r+1} \times 2^r, r \in \mathbb{N}$. Applying then the marching cubes algorithm [54], the 3D surface is extracted in the form of triangular meshes (vertex positions, normal vectors and connectivity). The coloring and texturing of each triangle of the surface is based on a weighted average between the cameras for which the specific part is not occluded. The weights estimation depends on the visibility angle between the camera and the respective area. Applying [8] in voxel grid resolutions with $r = 5$, $r = 6$, $r = 7$, we extract textured and colored triangular 3D mesh instances for all the mRGBD frames of the dataset in three (3) different resolutions. Color-per-vertex and textured 3D mesh instances are depicted in Fig. 10.

D. HUMAN4D BENCHMARKING SUBSETS

For benchmarking on HUMAN4D, we divide the dataset into two subsets, a single- (H4D1) and a two-person one (H4D2), in order to reduce the amount of data processing, as well as to evaluate samples that represent varying human poses. At the beginning of each sequence, the subjects were standing in T-Pose for calibration purposes. To that end, we decided to remove the first 100 frames of each sequence to avoid the collection of many similar poses (T-Pose) and to randomly sample 100 frames from the remaining part of each sequence, totaling 5600 and 1000 single-person and multi-person frames, respectively. Given that we benchmark HUMAN4D with pre-trained models or non data-driven encoders, both subsets, H4D1 and H4D2, are used as testing sets. The rest of the data can be considered as training and validation sets to allow the experimentation and development of new data-driven approaches on HUMAN4D. We benchmark HUMAN4D with respect to pose estimation and volumetric video compression by applying state-of-the-art approaches of the respective fields. In the

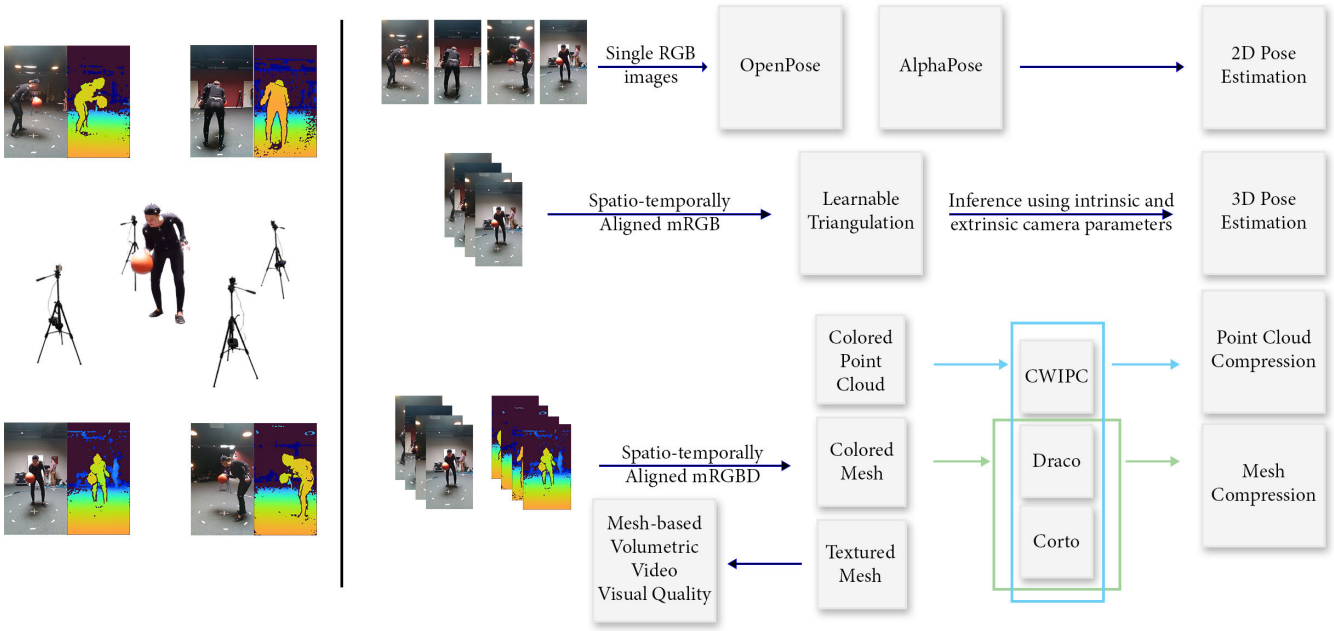

FIGURE 11. Overview of the benchmarking schema, given the spatio-temporally aligned mRGBD frames and ground-truth poses. Single-view RGB images are fed for 2D pose estimation. Multi-view RGB data are used for multi-view 3D pose estimation. Multi-RGBD frames are processed to produce point- and mesh-based volumetric video for 3D compression and visual quality benchmarking.

following sections (Sec. IV and V), we evaluate pre-trained models as well as 3D codecs for pose estimation and 3D compression respectively, on the benchmarking subsets of the dataset. An overview of the benchmarking flow and methodology we follow and present in the following sections is depicted in Fig. 11.

IV. POSE ESTIMATION

HUMAN4D enables research to human pose-related computer vision tasks by providing spatio-temporally aligned RGBD data from multiple views under a HW-SYNC setting, along with accurate 3D and 2D poses. Recent research efforts are devoted on various single- and multi-person pose estimation approaches, from single RGB in the wild [18], [57]–[59], depth [60], [61], multi-view RGB [23], [62] and multi-view RGBD [22], [63], among others. However, the selection criteria of the methods we benchmark are to be open-source and applicable to HUMAN4D, producing baseline results for our dataset. Finally, it is worth noting that the mRGBD frames of the evaluation set that go beyond the capabilities of the pre-trained models (for instance, several body parts out of at least one of the views) are excluded, preventing wrong and unfair evaluation with respect to the effectiveness of the methods.

A. SINGLE-VIEW 2D POSE ESTIMATION

Considering the 2D poses per view, we assess state-of-the-art methods for 2D pose estimation from color images. We apply the methods on the color views of all (4) RGBD cameras, extracting the overall error metrics per mRGB frame by averaging the errors per view.

Methods: We select 2 widely known 2D pose estimation methods, a bottom-up and a top-down one, to assess their effectiveness on HUMAN4D color images. Firstly, we select OpenPose by *Cao et al.* [21], a deep bottom-up pose estimation method that combines confidence maps with part affinity fields to predict multi-person 2D poses in real-time. For the evaluation of HUMAN4D, we used the latest version of the method as found to the official code repository.[3] Secondly, we evaluate AlphaPose, another data-driven approach proposed by Fang *et al.* [55]. AlphaPose constitutes a top-down, real-time 2D pose estimation method, that is continuously supported and updated over the last years. For the present experiments, we used the latest version of the method as found on the official repository of the authors.[4]

Finally, we also experimented with the official code of VNect,[5] by Mehta *et al.* [20], one of the first data-driven methods that approached 3D pose estimation from single RGB images, and A2j,[6] by Xiong *et al.* [60], for 3D pose estimation from single depth maps. However, the methods were not favorably applicable to our dataset, probably due to the differences between the characteristics of the training sets used to train the models and HUMAN4D. For A2j for

[3]https://github.com/CMU-Perceptual-Computing-Lab/openpose/tree/b5bffe18a8021f5f3ed98f19441b658647d9a8c3
[4]https://github.com/MVIG-SJTU/AlphaPose/tree/a22d3d6047b05be6ed94567c520d2a20d28d0407
[5]http://gvv.mpi-inf.mpg.de/projects/VNect
[6]https://github.com/zhangboshen/A2J/tree/60b45312c5009b2053d014510c08806c2c91e950

instance, the depth data used to train the body pose estimation model have been captured with Asus Xtion PRO, a structured-light depth sensor that provides depth maps of different resolution and depth noise in comparison with the stereo-based depth sensor from Intel, Intel RealSense D415. To this end, the results are not presentable, however the related tools for experimentation are available in the code repository of our dataset.[7]

Metrics: To measure the body joints localization accuracy, we measure mean Average Precision (mAP) for the common joints between the 2 methods and the ground truth annotations considering the Percentage of Correct Keypoints-head (*PCKh*) metric, as defined in [42]. *PCKh* constitutes a slight modification of Percentage of Correct Keypoints (*PCK*) [64], defining a matching threshold α as the percentage of the head segment length (from neck to head top), instead of the long edge of the bounding box that contains the subject, aiming to make the metric independent from specific body posture and articulation. To this end, a prediction for a frame f and a skeleton s is considered correct if its euclidean 2D distance error $\epsilon_{f,s}$ falls within a pixel circular region around the ground-truth keypoint with radius $r = \alpha L_{head}$, i.e.:

$$PCKh(f, s, j) = \begin{cases} 1, & \epsilon_{f,s}(j) \leq \alpha L_{head} \\ 0, & \epsilon_{f,s}(j) > \alpha L_{head} \end{cases} \quad (3)$$

$$AP_{PCKh}(f, s) = \frac{1}{\mathcal{J}_s} \sum_{j=1}^{\mathcal{J}_s} PCKh(f, s, j) \quad (4)$$

where L_{head} is the length of the head segment and α is the scalar that controls the relative threshold for correctness consideration.

Results: We separately present the results of the methods on H4D1 and H4D2 to better distinguish their effectiveness on single- and multi-person color data. At first, similarly to the outcomes on other public datasets, AlphaPose outperforms OpenPose showing higher accuracy both in single- and multi-person benchmarking sets of HUMAND. Nevertheless, even though both methods showcase lower accuracy on the multi-person data of H4D2, which is much more challenging due to the occlusions between the subjects, it is worth noting that the difference between the single- and multi-person results of OpenPose is low ($\sim$ 1.5%), while AlphaPose presents a higher drop of approximately 9%. Taking into account that the distance between the subjects and the sensors is short, from 1 to 2 meters, and in most of the two-person samples, there are severe occlusions for some of the sensors, we can probably assume that OpenPose, as a bottom-up approach behaves more robustly on occlusions, however AlphaPose, as a top-down approach, is more accurate but is strongly affected by occlusions. In order to provide extra information to the reader, along with the results on HUMAN4D, we also indicate the related outcomes of the methods to other datasets, i.e. MPII [42] and COCO [56] using *PCKh* with $\alpha = 0.5$, as presented in Table 3. Finally, a plot depicting the correlation between *PCKh* mAP against α threshold for both methods on both subsets, is illustrated in Fig. 12.

TABLE 3. 2D pose estimation results of OpenPose [21] and AlphaPose [55] with $AP_{PCKh\text{-}0.5}$.

mAP (%)	MPII [42]	COCO [56]	H4D1	H4D2
Cao *et al.* OpenPose [21]	72.50	64.20	**70.02**	**68.48**
Fang *et al.* AlphaPose [55]	82.10	71.00	**82.95**	**73.94**

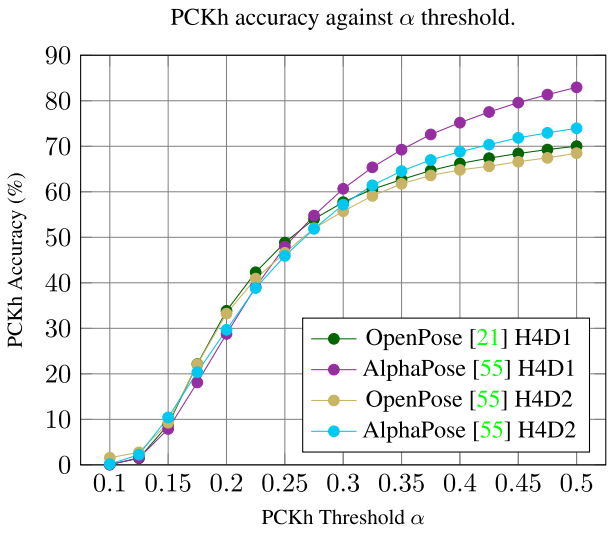

FIGURE 12. OpenPose [21] and AlphaPose [55] applied on the 4 views of the RGBD cameras on H4D1 and H4D2, extracting the overall error metrics per mRGB frame by averaging the errors per joint.

B. MULTI-VIEW 3D POSE ESTIMATION

Subsequently, we evaluate multi-view 3D pose estimation on HUMAN4D, exploiting the multi-view color images along with the respective intrinsic and extrinsic camera parameters and using HUMAN4D 3D poses as ground truth.

Methods: We choose a recent state-of-the-art method proposed by Iskakov *et al.* [23], which constitutes a novel solution for multi-view single-person 3D human pose estimation based on a learnable triangulation (LT) technique, combining 3D information from multiple spatio-temporally aligned 2D color views. In particular, $LT_{(alg.)}$ [23] is a top-down 3D pose estimation method based on end-to-end differentiable algebraic triangulation with an addition of confidence weights estimated from the input images. We ran the experiments only on the HD41 benchmarking subset of the dataset since the method estimates single-person 3D poses, using the latest version of the code published by the authors.[8]

Metrics: With respect to the metrics, we use the Mean Per Joint Position (MPJP) [20] and Root Mean Squared Per Joint Position (RMSPJP) error metrics, which both are influenced by large outliers, however the latter better incorporates the

[7]https://github.com/tofis/human4d_dataset

[8]https://github.com/karfly/learnable-triangulation-pytorch

TABLE 4. Single-person pose estimation results on H4D1 and CMU [39].

Datasets	CMU	HUMAN4D (H4D1)			
Metrics	**MPJP** (cm)	**MPJP** (cm)	**RMSPJP** (cm)	**mAP** ($\mathbf{PCK}_{\alpha_{3D}=10cm}$)	**mAP** ($\mathbf{PCK}_{\alpha_{3D}=12.5cm}$)
Iskakov *et. al* LT (alg.) [23]	2.13	8.42	9.56	80.26%	86.52%

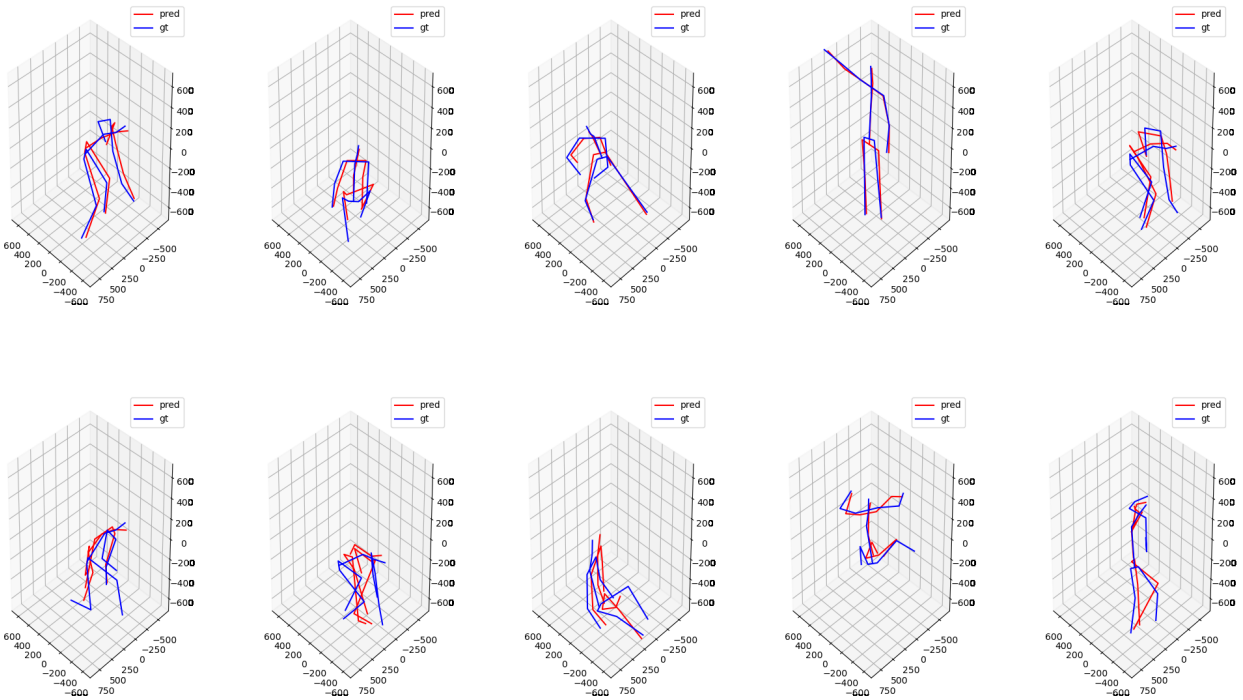

FIGURE 13. Qualitative results of learnable triangulation (alg.) proposed by Iskakov *et al.* [23]. The top and bottom rows depict success and failure cases, respectively. **Blue** and **red** colors correspond to **ground truth** and **predicted** poses.

variance of the estimates and their bias. For a frame f and a skeleton s, MPJP and RMSPJP are computed as:

$$\epsilon_{f,s}(j) = ||\hat{x}_{f,s}(j) - x_{f,s}(j))||_2 \tag{5}$$

$$\mathcal{E}_{MPJP}(f, s) = \frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s}\epsilon_{f,s}(j) \tag{6}$$

$$\mathcal{E}_{RMSPJP}(f, s) = \sqrt{\frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s}\epsilon_{f,s}^2(j)} \tag{7}$$

where $\mathcal{J}_s$ is the total number of joints of skeleton s. Finally, we also use mean AP with 3D PCK metric [65] per joint, where an estimate is considered correct when the 3D euclidean distance error, i.e. $\epsilon_{f,s}(j)$, is less than a distance threshold α_{3D}, as:

$$PCK_{3D}(f, s, j) = \begin{cases} 1, & \epsilon_{f,s}(j) \leq \alpha_{3D} \\ 0, & \epsilon_{f,s}(j) > \alpha_{3D} \end{cases} \tag{8}$$

$$AP_{PCK_{3D}}(f, s) = \frac{1}{\mathcal{J}_s}\sum_{j=1}^{\mathcal{J}_s} PCK_{3D}(f, s, j) \tag{9}$$

for a frame f and skeleton s, correspondingly.

Results: Classic triangulation algorithms assume that the 2D point coordinates from each view equally contribute to the triangulation 3D point coordinates estimation. The major advantage of the LT approach is that the contribution of the 2D joint positions that cannot be estimated reliably (e.g. due to joint occlusions) to the final triangulation outcome, is controlled by a neural network. In particular, learnable weights have been added to the coefficients of the matrix corresponding to different views. A limitation of the LT approach is that it fails when some of the body parts are out of the field of view of the cameras, leading to erroneous estimates. Another limitation is that LT approach supports only single-person 3D pose estimation and for that reason it was applied only on H4D1. Quantitative results of the method on HUMAN4D, complemented with results on CMU [39] dataset, are reported in Table 4. Fig. 14 illustrates the correlation between the mAP against α_{3D} threshold on HUMAN4D. Qualitative results regarding the predicted 3D poses against ground-truth on HUMAN4D are illustrated in Fig. 13, where $\text{LT}_{(\text{alg.})}$ seems accurate in "clean" poses where self-occlusions are limited (success cases on top rows), while the accuracy is limited in the presence of self-occlusions (failure cases on bottom rows).

V. VOLUMETRIC VIDEO

Beyond pose estimation, we benchmark a set of state-of-the-art static 3D codecs, in the context of a live streaming

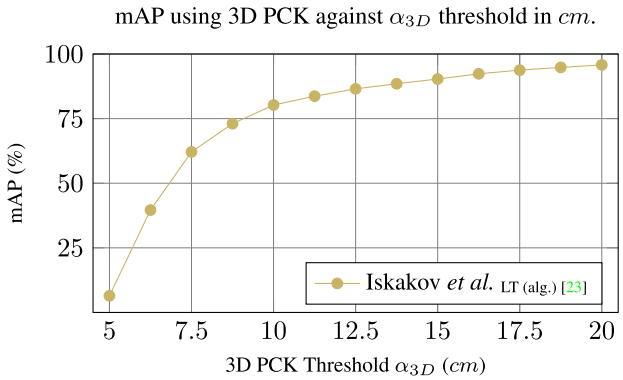

FIGURE 14. Benchmarking of Algebraic Learnable Triangulation [23] on H4D1 using total 3D PCK results in different α_{3D} threshold values in *cm*.

scenario. Moreover, we assess the visual quality of textured 3D mesh instances to demonstrate the positive correlation between the objective visual quality and the FFT voxel-grid resolution.

A. VOLUMETRIC VIDEO COMPRESSION

Compression of volumetric data produced in a real-time manner is thought to be a key enabler of a wide variety of applications, such as XR teleconference, real-time dense surface mapping in AR devices and free-viewpoint videos. A key contribution of HUMAN4D is that it enables future benchmarking in static and temporal volumetric video compression, by offering a large dataset of samples and sequences of point- and mesh-based volumetric data. In contrast with motion pictures where solutions are mature and proven, real-time varying geometry coding is still an open challenge frequently cured utilizing only intra-frame coding, ignoring temporal relations between volumes of consecutive frames. Such an endeavour is presented in [66] by Doumanoglou *et al.* In a similar manner, for the purpose of this work, the codecs are tested in various profiles, aiming at specific bit-rates, using appropriate metrics on HUMAN4D point- and mesh-based volumetric data cues. To be coherent, we define common codec profiles both for H4D1 and H4D2 dataset subsets. A matching procedure between different codecs for the same target bit-rate was adopted, defining the acceptable deviation margin between target and achieved bit-rate to be $\pm 10\%$.

1) MESH-BASED VOLUMETRIC VIDEO COMPRESSION

Initially, we benchmark 3D codecs on mesh-based volumetric data using the benchmarking subsets of meshes reconstructed in three different voxel-grid resolutions (i.e. $r = \{5, 6, 7\}$) applying the real-time 3D reconstruction method by Alexiadis *et al.*, as reported in Section III-C3.

Codecs: We employ Corto [67] and Draco [68], two 3D codecs particularly chosen due to their high quality real-time performance. Targeting specific bit-rates for real-time mesh-based volumetric video transmission, we constructed a series of compression profiles with varying compression level, quantization parameter per attribute and different compression methods for specific attributes. HUMAN4D mesh-based compression benchmarking focuses on three different per-vertex attributes: *geometry* and *normals* represented in floating points and *color* in unsigned integers.

Corto codec [67] configuration consists of four parameters. One quantization value for each of the mesh attributes, i.e. Geometry (GQ), Normal (NQ) and Color (CQ) Quantization bits, and one switch to denote the normal prediction method. We select between two different normal prediction methods, the Normals Quantized Coding (NQC) and the Normals Delta Coding (NDC). In the former, we store the differences between the normals estimated from the quantized geometry and the quantized actual normals, using an octahedron projection representation [69]. In the latter, the quantized normals in the octahedron projection representation are solely delta coded, with respect to a neighboring quantized normal belonging to a quad incident to the normal's vertex.

Regarding the Draco codec [68], the configurable parameters are the compression level (CL) which adjusts the compression speed versus the size mixture, the geometry quantization bits (GQ), the normals quantization (NQ) and the color quantization bits (CQ). Contrary to Corto, Draco does not expose any normal manipulation option to adjust.

Beyond these conventional open-source codecs, novel 3D and 4D data compression approaches have appeared, such as the one proposed by Tang *et al.* [36]. This method constitutes a novel block-based 3D compression model, being the first deep 3D compression method that can train end-to-end with entropy coding, lossless compression of the surface topology, exhibiting a novel block-based texture parametrization that inherently promotes temporal consistency without tracking and the necessity of the UV coordinates compression. This codec achieves superior results in comparison to conventional 3D codecs, such as Draco and Corto, in regards with the rate-distortion (RD) balance. Specifically, it is deemed to achieve on average 66% lower bit-rate for the same level of distortion in 4D data. For the purpose of this work, we did not benchmark this particular codec since it is not currently open-source.

Metrics: With respect to the metrics, we use *RMS*, *HausdorffAbs* and *HausdorffRel* metrics to compare the compressed and raw mesh-based representations. For the extraction of *RMS* and *Hausdorff* distance, we exploit a tool implemented based on [70]. This tool provides numerical metrics for the similarity of source and target triangle or quadrilateral meshes. It is worth mentioning that, for the same pair, swapping between the source and target meshes can lead to different numerical values, thus as usual for these metrics in the literature, we define the correct value to be the maximum of these two, for all metrics.

Hausdorff distance metric is used in two variations. *HausdorffAbs metric* is defined as the maximum value of all the uniformly minimum sampled distances across all points of the source surface to the target surface. *HausdorffRel metric* is a variation of *HausdorffAbs metric* which tackles

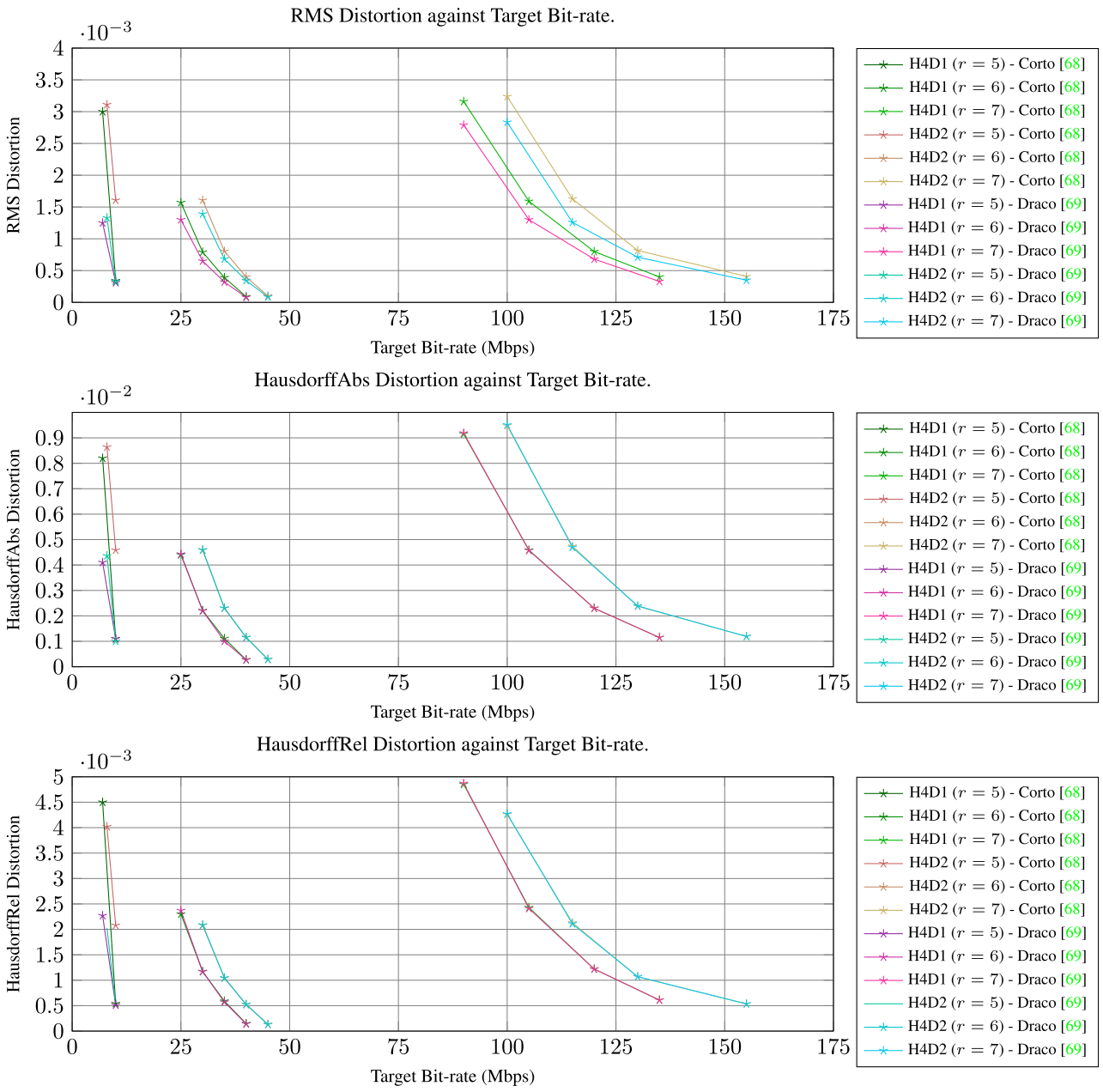

FIGURE 15. 3D Mesh *RMS*, *HausdorffAbs* and *HausdorffRel* distortions vs target bit-rates on H4D1 and H4D2.

the comparison of surfaces with different scales. For the *RMS* calculation, we need to have a set of minimum distances between two surfaces, the mean distance E_m can be calculated by:

$$E_m(S, S') = \frac{1}{|S|} \int_S d(p, S') dS \tag{10}$$

where $|S|$ denotes the area of S. Using the mean distance formula, the *root mean square error* is defined by:

$$RMS_{S \to S'} = \sqrt{\frac{1}{|S|} \int_S d(p, S') dS}. \tag{11}$$

Results: For a fair comparison between the codecs, we choose to employ a testing scheme based on rate-distortion terms. In that direction, we keep the bit-rates steady for the pairs and evaluate the corresponding distortion introduced by each codec. As it can be seen in Fig. 15, Draco consistently outperforms Corto, in terms of distortion induced for any tested bit-rate. The profiles used for the benchmarking are depicted in Table 5.

Having tested the same codec profiles both for single and multi-person subsets of the HUMAN4D dataset, we noticed that the bit-rates achieved by both codecs on the multi-person subset are slightly greater than those on the single-person

TABLE 5. Draco [68] and Corto [67] Codec configurations used to achieve the targeted bit-rates for the voxel-grid resolutions of the reconstruced 3D mesh instances, i.e. for $r = 5$, $r = 6$ and $r = 7$.

Voxel-Grid Resolution	Target Bit-rate (Mbps) H4D1/2	Draco [69] Codec Configuration	Corto [68] Codec Configuration
$r = 5$	7/8	(CL 6, GQ 8, NQ 8, CQ 5)	(GQ 8, NQ 8, CQ 5, NQC)
	10	(CL 4, GQ 10, NQ 10, CQ 5)	(GQ 11, NQ 11, CQ 5, NQC)
$r = 6$	25/30	(CL 6, GQ 8, NQ 8, CQ 5)	(GQ 9, NQ 9, CQ 5, NDC)
	30/35	(CL 5, GQ 9, NQ 9, CQ 5)	(GQ 10, NQ 10, CQ 5, NDC)
	35/40	(CL 6, GQ 10, NQ 10, CQ 5)	(GQ 11, NQ 11, CQ 5, NQC)
	40/45	(CL 6, GQ 12, NQ 12, CQ 5)	(GQ 13, NQ 13, CQ 5, NQC)
$r = 7$	90/100	(CL 2, GQ 7, NQ 7, CQ 5)	(GQ 8, NQ 8, CQ 5, NDC)
	105/115	(CL 2, GQ 8, NQ 8, CQ 5)	(GQ 9, NQ 9, CQ 5, NDC)
	120/130	(CL 4, GQ 9, NQ 9, CQ 5)	(GQ 10, NQ 10, CQ 5, NDC)
	135/155	(CL 5, GQ 10, NQ 10, CQ 5)	(GQ 11, NQ 11, CQ 5, NQC)

TABLE 6. CWIPC [33], Draco [68] and Corto [67] Codec configurations used to achieve the targeted bit-rates for the reconstructed (R) and sampled (S) point-clouds.

PC (R/S)	Target Bit-rate (Mbps)	CWIPC [33] Codec Configuration	Draco [69] Codec Configuration	Corto [68] Codec Configuration
R	2/3	(OD 6, JPEGQP 75)	-	-
	4/7	(OD 7, JPEGQP 65)	-	-
	7	(OD 8, JPEGQP 75)	(CL 10, CQ 5, CQ 8)	(GQ 10, CQ 6)
	9/15	(OD 9, JPEGQP 65)	(CL 10, CQ 6, CQ 8)	(GQ 11, CQ 6)
S	4/8	(OD 7, JPEGQP 75)	-	-
	15/25	(OD 8, JPEGQP 85)	-	-
	40/50	(OD 9, JPEGQP 85)	(CL 3, GQ 5, CQ 8)	(GQ 10, CQ 6)
	70/90	(OD 10, JPEGQP 85)	(CL 10, GQ 8, CQ 8)	(GQ 11, CQ 6)

one. That is probably due to the fact that the additional information induced in the form of the second subject, leads to larger surfaces that, despite using the same voxel-grid areas and resolutions, results in more challenging 3D surfaces to compress, in regards with elements count and connectivity information.

2) POINT-BASED VOLUMETRIC VIDEO COMPRESSION

To benchmark point cloud compression, beyond the reconstruction of the raw point-cloud instances from the mRGBD samples described in Section III-C3, we also use another point-cloud reconstruction approach. The raw point-cloud instances typically contain $\sim 25,000$ points per frame for the single-subject sequences and $\sim 40,000$ points for the two-subject ones. This alternative reconstruction approach allows us to create denser point clouds by sampling points from the surface of the high resolution meshes (i.e. using voxel-grid resolution with $r = 7$). Points are sampled from the mesh surface with a probability proportional to the area of the underlying mesh faces using Point Cloud Library (PCL) [71]. We set the algorithm to generate point cloud instances containing 300, 000 points per frame.

Codecs: To benchmark the performance of point cloud compression, we perform a rate-distortion analysis for the codecs Draco, Corto and CWIPC, the MPEG anchor codec proposed in [33] and evaluated in [72]. CWIPC is parameterizable with respect to the Octree Depth (OD) and JPEG Quantization Parameter (JPEGQP). We select to perform the analysis on 4 target bit-rates. Note that, for all codecs we first identified the compression parameters that achieve the target bit-rates within a 10% tolerance. Details on these profiles are listed in Table 6.

Metrics: To measure the distortions introduced by compression to the point-cloud samples, we used standard, well established, full reference metrics, as released by the standards body MPEG [73], [74]. More specifically, we measure Peak Signal-to-Noise Ratio (PSNR) using the maximum of the nearest neighbor euclidean distances amongst all points in the reference point cloud as the peak value v_p by:

$$PSNR = 10 \log(\frac{v_p^2}{MSE}) \tag{12}$$

The same process is then applied to the point cloud colors at each of the corresponding points between the decoded and the groundtruth point clouds. Metrics are collected utilizing the MPEG PCC-DMETRIC tool [75][9] to calculate these distortions for each frame in the dataset.

Results: Analyzing the experimental results, CWIPC codec achieves lower geometry distortions for the same bit-rate in comparison with Draco and Corto, while in higher bit-rates, all the benchmarked codecs showcase similar efficiency. CWIPC exploits octree occupancy to encode geometry positions, thus is able to retain more points from the original point cloud. Details with respect to point-cloud compression benchmarking are illustrated in Fig. 16, while the codec profiles used for the experiments are listed in Table 6. For the sake of clarity, we summarize the abbreviations of codec configuration parameters in Table 7.

[9]http://mpegx.int-evry.fr/software/MPEG/PCC/mpeg-pcc-dmetric

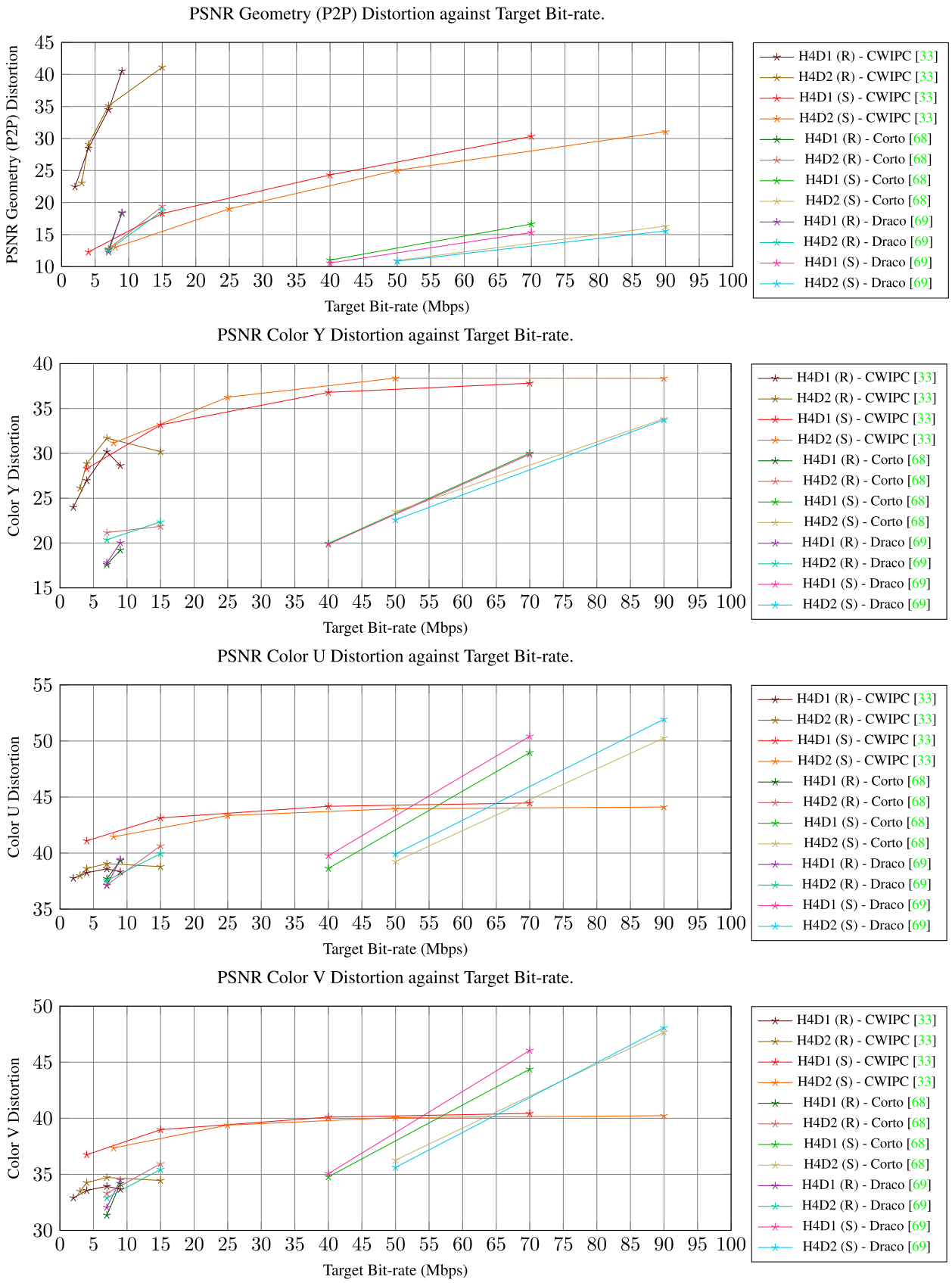

FIGURE 16. Point-cloud *PSNR Geometry* and *Color YUV* distortions vs target bit-rates on H4D1 and H4D2.

TABLE 7. Abbreviations.

Codecs	Parameter	Abbreviation
Draco	Compression Level	CL
Draco/Corto	Normal Quantization Bits	NQ
Draco/Corto	Geometry Quantization Bits	GQ
Draco/Corto	Color Quantization Bits	CQ
Corto	Normals Quantized Coding	NQC
Corto	Normals Delta Coding	NDC
CWIPC	Octree Depth	OD
CWIPC	JPEG Quantization Parameter	JPEGQP

B. MESH-BASED VOLUMETRIC VIDEO VISUAL QUALITY

In this section, we assess the visual quality of HUMAN4D textured 3D mesh instances between the three different resolutions of the underlying voxel-grid. The aim is to demonstrate the positive correlation between the objective visual quality and the utilized voxel-grid resolution used to reconstruct the mesh-based volumetric data.

As mentioned in Section III-C3, the reconstruction of the mesh-based volumetric data is achieved by applying the real-time method proposed by Alexiadis *et al.* [8], parameterized in three different voxel-grid resolutions to produce watertight textured 3D mesh instances of varying vertex and face counts. Higher resolution grids lead to meshes of higher element count that are, per se, expected to capture more photorealistically and precisely the observed subjects.

Apart from the self-evident impact of higher resolution sampling on the reconstructed hull's spatial fidelity, additional benefits may arise with regard to the accurate colorization of its surface. To showcase and quantify this effect, we firstly project the examined mesh on its respective RGB images and sample the color of its fragments based on a weighted contribution of the corresponding pixels. Then, we render the mesh from the exact same viewpoints that the aforementioned images were captured and compare the synthesized images to their respective silhouette-cropped textures, using conventional image quality metrics.

We conduct the assessment separately to H4D1 and H4D2 benchmarking subsets. The former, consisting of 4 subjects with 14 sequences each, and each of these sequences with 100 sampled mRGBD frames, reconstructed in 3 voxel-grid resolutions (i.e. $r = \{5, 6, 7\}$) and rendered from 4 viewpoints, results in a total of $67,200$ rendered views of $16,800$ mesh instances. Similarly, the latter includes 2 couples, with 5 sequences of 100 frames each, reconstructed in the same 3 voxel-grid resolutions and rendered from corresponding viewpoints, giving a total of $12,000$ views of $3,000$ 3D meshes.

Metrics: For the visual quality assessment, we opted to use *Peak Signal-to-Noise Ratio (PSNR)* (Eq. 12) and *Structural Similarity Index (SSIM)* as metrics to objectively quantify the photometric and photorealistic consistency between the captured, raw color (RGB) views and the mesh-based 4D representations in the various voxel-grid resolutions on the rendered views' quality.

SSIM is a full-reference metric conceived as an improvement over the traditional *PSNR* and *MSE-family* metrics and is widely referenced in the video and photography industry as it is believed to capture better the human perception of visual quality. Instead of decomposing the input signals and then estimating absolute errors, as in the case of MSE-like metrics, *SSIM* incorporates into its calculations the fact that images are inherently highly structured and thus their topology and the relations that arise between their elements, due to that fact, should not be ignored. Luminance Masking and Contrast Masking are two well-known visual perception phenomena that are taken into account during the process of obtaining *SSIM* measurements. The former is about the low visibility of distortions in bright regions, while the latter is about the masking of distortions in highly textured, non-smooth, areas of an image.

The *SSIM* formula is composed of three individual measurements of "structural similarity", luminance l, contrast c and structure s between two windows x and y of similar size. The individual comparison formulas are:

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{13}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{14}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{15}$$

with μ_x the average of x, μ_y the average of y, σ_x^2 the variance of x, σ_y^2 the variance of y, σ_{xy} the covariance of x and y, $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$, $c_3 = (c_2/2)$ are three variables to stabilize the division with weak denominator, L the dynamic range of the pixel values and $k_1 = 0.01$, $k_2 = 0.03$ by default. *SSIM* is then a weighted combination of these comparative measures:

$$SSIM(x, y) = [l(x, y)^\alpha \cdot c(x, y)^\beta \cdot s(x, y)^\gamma] \tag{16}$$

where $\alpha, \beta, \gamma > 0$ are parameters used to adjust the relative importance of the three components. More on the SSIM and its development can be found in [76].

Results: As can be seen in Tables 8 and 9, the experiments conducted, validate the claim that increments of a textured mesh voxel-grid resolution lead to increases in its objective visual quality. Both for single- and multi-person evaluation sets, *PSNR* increases in par with mesh resolution. From $r = 5$

TABLE 8. Single-person PSNR and SSIM.

	PSNR			SSIM		
Subject	r=5	r=6	r=7	r=5	r=6	r=7
S1	36.18	36.59	**36.70**	0.98598	0.98685	**0.98707**
S2	34.51	34.84	**34.89**	0.98320	0.98389	**0.98395**
S3	33.48	33.71	**33.73**	0.98235	**0.98278**	0.98270
S4	33.36	33.54	**33.55**	0.98262	**0.98302**	0.98293
Average	34.38	34.67	**34.72**	0.98354	0.98413	**0.98416**

(a) $r = 5$ (b) $r = 6$ (c) $r = 7$ (d) *RGB*

(e) $r = 5$ (f) $r = 6$ (g) $r = 7$ (h) *RGB*

FIGURE 17. Textured mesh-based volumetric samples from H4D1 and H4D2 rendered in the 3 different voxel-grid resolutions along with the corresponding RGB images from the same viewpoint.

TABLE 9. Multi-person PSNR and SSIM.

	PSNR			SSIM		
Subjects	r=5	r=6	r=7	r=5	r=6	r=7
S1 & S2	32.59	33.02	**33.26**	0.97513	0.97634	**0.97720**
S3 & S4	38.20	38.37	**38.41**	0.98346	0.98432	**0.98488**
Average	35.39	35.70	**35.84**	0.97930	0.98033	**0.98104**

to $r = 6$ the increase is more pronounced, while from $r = 6$ to $r = 7$, it seems to diminish, indicating that a further increase in 3D mesh voxel-grid resolution may be futile, at least as regards the texture fidelity in terms of *PSNR*. The *SSIM* case generally follows the same trend, with the exception of the *S3* and *S4* subjects from the single-person subset, where post $r = 6$ increase in resolution does not seem to further improve the *SSIM* of the textures. In these cases, the $r = 6$ and $r = 7$ *SSIM* values are approximately equal, exhibiting a difference of less than 10^{-4}.

In Fig. 17, volumetric samples from the single- and multi-person subsets are illustrated, rendered in the 3 different voxel-grid resolutions along with the corresponding RGB images from the same viewpoint. The increase of texture quality we want to highlight in these views is most apparent in the eyes area of the multi-person renderings. As can be seen, for $r = 5$ the right eye of the male subject is blurry and barely visible. As the voxel-grid resolution increases, the eye gets crisper and better defined. Such behaviour can be noticed in other areas of the volumetric data as well.

In a nutshell, experimental results indicate that the increase of 3D mesh voxel-grid resolution indeed leads to objective quality increase, though with diminishing returns. This latter observation, together with the near real-time capabilities of the mesh-based volumetric reconstruction pipeline for $r = 6$ and the decreased bandwidth needs it requires when compared with the $r = 7$ case, makes $r = 6$ voxel-grid resolution the most sensible choice for a volumetric live-streaming setup.

VI. DISCUSSION

We created HUMAN4D to provide the research community with a public resource that fills identified gaps in publicly available human-centric 4D datasets, consisting of motion capture and HW-SYNCed volumetric data. In the flood of

recent literature, a plethora of algorithms and deep models focus on 3D pose estimation, however, only a few methods approach the task with the use of multi-view depth and volumetric data. That is probably due to the complexity and time-consuming setup of multi-view capturing settings as well as the lack of spatio-temporally aligned multi-view depth maps with ground-truth data. To this end, we aim to enable research on that direction encouraging the computer vision community to develop and experiment with new 3D pose estimation approaches on HUMAN4D by providing HW-SYNCed depth and volumetric data along with ultra-accurate ground-truth 3D poses for supervision and evaluation. With regards to volumetric data, volumetric video is an emerging immersive medium, being unique due to its fully three-dimensional nature and its capability to enable six degrees of freedom (6DoF) spectating when used in 4D environments. HUMAN4D has been created on the principle to provide spatio-temporally aligned mRGBD data captured to produce point- and mesh-based volumetric videos, reconstructed and compressed respecting online encoding and steady bit-rates. On top of that, in most public datasets, the temporal misalignment between the multiple color and depth streams adds extra noise to the already noisy depth and color data, reducing the quality of the volumetric video. In HUMAN4D, this noise is absent due to the high synchronization precision (HW-SYNC).

VII. CONCLUSION

In this paper we introduced HUMAN4D, a new multimodal human-centric 4D dataset containing a large corpus with more than 50K samples from daily, physical and social activities of annotated spatio-temporally aligned multi-view RGBD, volumetric and motion capture data along with audio recordings. To the best of our knowledge, HUMAN4D is the first dataset that provides HW-SYNCed mRGBD frames with the use of recent consumer-grade depth sensing devices. We also provide evaluation benchmarks based on discriminative pose estimation and volumetric data compression methods. We make all the data[10] and code[11] available online, including the respective synchronization, calibration and camera parameters, along with data loaders and other processing, visualization and evaluation tools, for academic use and further research. In that scope, the authors commit to continuously maintain the dataset for the community by adding new tools, baselines and captures. Despite the continuous maintenance of the dataset, benchmarking subsets will remain constant to allow the assessment and comparison between new state-of-the-art methods on the same datasets. We believe that HUMAN4D and its associated tools will stimulate further research in computer vision and data driven approaches, enabling research on human pose estimation, real-time volumetric video reconstruction and compression, with the use of consumer-grade RGBD cameras sensors.

[10]http://dx.doi.org/10.21227/xjzb-4y45

[11]https://github.com/tofis/human4d_dataset

ACKNOWLEDGMENT

The authors gratefully appreciate the work conducted by the team of the Artanim Foundation Motion Capture Studio, providing high quality motion capture and 3D scanning services. They also want to give special thanks to Sylvain Chagué and Valérie Juillard, members of Artanim team, for scanning, post-processing and rigging of the 3D character and for post-processing and retargeting of the animations, respectively.

Author Contributions - A.C: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Supervision; L.S: Methodology, Software, Validation, Formal Analysis, Investigation; P.B: Methodology, Software, Formal Analysis, Investigation, Data Curation; P.D: Software, Validation, Formal Analysis, Investigation; N.Z: Methodology, Software; S.S: Methodology, Investigation, Data Curation; B.K: Methodology, Data Curation, Resources; C.C: Methodology, Data Curation, Resources; P.C: Methodology; D.Z: Supervision; S.K: Supervision; P.D: Supervision, Funding Acquisition.

REFERENCES

[1] A. Mukherjee, A. K. Panja, and N. Dey, *A Beginner's Guide to Data Agglomeration and Intelligent Sensing*. New York, NY, USA: Academic, 2020.

[2] S. N. Gunkel, H. M. Stokking, M. J. Prins, N. van der Stap, F. B. T. Haar, and A. O. Niamut, "Virtual reality conferencing: Multi-user immersive VR experiences on the Web," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 498–501.

[3] H. Lee, T. H. Jung, M. C. tom Dieck, and N. Chung, "Experiencing immersive virtual reality in museums," *Inf. Manage.*, vol. 57, no. 5, Jul. 2020, Art. no. 103229.

[4] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black, "Dynamic FAUST: Registering human bodies in motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6233–6242.

[5] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. Black, "AMASS: Archive of motion capture as surface shapes," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5442–5451. [Online]. Available: https://amass.is.tue.mpg.de

[6] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–16, Nov. 2015.

[7] R. A. Newcombe, D. Fox, and S. M. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 343–352.

[8] 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 Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 798–813, Apr. 2017.

[9] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi, "Fusion4D: Real-time performance capture of challenging scenes," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–13, Jul. 2016.

[10] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, and V. Tankovich, "Holoportation: Virtual 3D teleportation in real-time," in *Proc. 29th Annu. Symp. User Interface Softw. Technol.*, 2016, pp. 741–754.

[11] A. S. Jackson, C. Manafas, and G. Tzimiropoulos, "3D human body reconstruction from a single image via volumetric regression," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–14. [Online]. Available: https://openaccess.thecvf.com/content_ECCVW_2018/papers/11132/Jackson_3D_Human_Body_Reconstruction_from_a_Single_Image_via_Volumetric_ECCVW_2018_paper.pdf

[12] 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 Trans. Graph.*, vol. 36, no. 4, p. 1, 2017.

[13] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll, "Video based reconstruction of 3D people models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8387–8397.

[14] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian, and D. Tang, "The relightables: Volumetric performance capture of humans with realistic relighting," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–19, 2019.
[15] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.
[16] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.
[17] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.
[18] R. A. Guler and I. Kokkinos, "Holopose: Holistic 3D human reconstruction in-the-wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 10884–10894.
[19] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 2252–2261.
[20] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt, "Vnect: Real-time 3D human pose estimation with a single rgb camera," *ACM Trans. Graph. (TOG)*, vol. 36, no. 4, pp. 1–14, 2017.
[21] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," 2018, *arXiv:1812.08008*. [Online]. Available: http://arxiv.org/abs/1812.08008
[22] M. Carraro, M. Munaro, J. Burke, and E. Menegatti, "Real-time markerless multi-person 3D pose estimation in RGB-depth camera networks," in *Proc. Int. Conf. Intell. Auton. Syst.* Springer, 2018, pp. 534–545.
[23] K. Iskakov, E. Burkov, V. Lempitsky, and Y. Malkov, "Learnable triangulation of human pose," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7718–7727.
[24] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 4342–4351.
[25] A. Chatzitofis, D. Zarpalas, S. Kollias, and P. Daras, "DeepMoCap: Deep optical motion capture using multiple depth sensors and retro-reflectors," *Sensors*, vol. 19, no. 2, p. 282, 2019.
[26] S. Tripathi, S. Ranade, A. Tyagi, and A. Agrawal, "PoseNet3D: Unsupervised 3D human shape and pose estimation," 2020, *arXiv:2003.03473*. [Online]. Available: http://arxiv.org/abs/2003.03473
[27] D. S. Alexiadis and P. Daras, "Quaternionic signal processing techniques for automatic evaluation of dance performances from MoCap data," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1391–1406, Aug. 2014.
[28] F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, "Motion analysis: Action detection, recognition and evaluation based on motion capture data," *Pattern Recognit.*, vol. 76, pp. 612–622, Apr. 2018.
[29] G. Kordelas, J. P.-M. Agapito, J. V. Hernandez, and P. Daras, "State-of-the-art algorithms for complete 3D model reconstruction," in *Proc. Engage Summer School*, Zermatt, Switzerland, vol. 1315, 2010, p. 115.
[30] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "LiveCap: Real-time human performance capture from monocular video," *ACM Trans. Graph.*, vol. 38, no. 2, pp. 1–17, Apr. 2019.
[31] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, and C. Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5052–5063.
[32] A. Doumanoglou, D. S. Alexiadis, D. Zarpalas, and P. Daras, "Toward real-time and efficient compression of human time-varying meshes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 12, pp. 2099–2116, Dec. 2014.
[33] R. Mekuria, K. Blom, and P. Cesar, "Design, implementation, and evaluation of a point cloud codec for tele-immersive video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 828–842, Apr. 2017.
[34] M. Quach, G. Valenzise, and F. Dufaux, "Learning convolutional transforms for lossy point cloud geometry compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4320–4324.
[35] J. Wang, H. Zhu, Z. Ma, T. Chen, H. Liu, and Q. Shen, "Learned point cloud geometry compression," 2019, *arXiv:1909.12037*. [Online]. Available: http://arxiv.org/abs/1909.12037
[36] D. Tang, S. Singh, P. A. Chou, C. Hane, M. Dou, S. Fanello, J. Taylor, P. Davidson, O. G. Guleryuz, Y. Zhang, and S. Izadi, "Deep implicit volume compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1293–1303.
[37] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Proc. IEEE Workshop Appl. Comput. Vis. (WACV)*, Jan. 2013, pp. 53–60.
[38] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[39] 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 *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3334–3342.
[40] Z. Yu, J. Shin Yoon, I. Kyu Lee, P. Venkatesh, J. Park, J. Yu, and H. S. Park, "HUMBI: A large multiview dataset of human body expressions," 2018, *arXiv:1812.00281*. [Online]. Available: http://arxiv.org/abs/1812.00281
[41] L. Sigal, A. O. Balan, and M. J. Black, "HUMANEVA: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, p. 4, 2010.
[42] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3686–3693.
[43] *PhaseSpace*. [Online]. Available: http://www.phasespace.com
[44] Google Inc. *Turbo Colormap*. Accessed: Aug. 20, 2019. [Online]. Available: https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html
[45] A. Grunnet-Jepsen, P. Winer, A. Takagi, J. Sweetser, K. Zhao, T. Khuong, D. Nie, and J. Woodfill, "Using the realsense d4xx depth sensors in multi-camera configurations," Intel Corp., Santa Monica, CA, USA, Tech. Rep. Rev 0.4, 2018.
[46] V. Sterzentsenko, A. Karakottas, A. Papachristou, N. Zioulis, A. Doumanoglou, D. Zarpalas, and P. Daras, "A low-cost, flexible and portable volumetric capturing system," in *Proc. 14th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, 2018, pp. 200–207.
[47] A. Papachristou, N. Zioulis, D. Zarpalas, and P. Daras, "Markerless structure-based multi-sensor calibration for free viewpoint video capture," Vaclav Skala Union Agency, Pilsen, Czech Republic, Tech. Rep., 2018. [Online]. Available: http://wscg.zcu.cz/WSCG2018/2018-papers/!!_CSRN-2801-10.pdf, doi: 10.24132/CSRN.2018.2801.10.
[48] (2019). *AgiSoft Metashape Professional (Version 1.5.2) (Software)*. Accessed: Mar. 2, 2020. [Online]. Available: http://www.agisoft.com/downloads/installer/
[49] CHDK Development Team. *Canon Hack Development Kit*. Accessed: Feb. 13, 2020.[Online]. Available: http://chdk.wikia.com/wiki/CHDK
[50] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.
[51] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 631–648.
[52] M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, and J. Huang, "Uncertainty-aware audiovisual activity recognition using deep Bayesian variational inference," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6301–6310.
[53] M. Kazhdan, "Reconstruction of solid models from oriented point sets," in *Proc. 3rd Eurographics Symp. Geometry Process.*, 2005, p. 73-es.
[54] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.
[55] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. ICCV*, 2017, pp. 2334–2343.
[56] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 740–755.
[57] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 483–499.
[58] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 468–475.
[59] M. Li, Z. Zhou, J. Li, and X. Liu, "Bottom-up pose estimation of multiple person with bounding box constraint," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 115–120.

[60] F. Xiong, B. Zhang, Y. Xiao, Z. Cao, T. Yu, J. T. Zhou, and J. Yuan, "A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 793–802.
[61] P. Szczuko, "Deep neural networks for human pose estimation from a very low resolution depth image," *Multimedia Tools Appl.*, vol. 78, no. 20, pp. 29357–29377, Oct. 2019.
[62] M. Kocabas, S. Karagoz, and E. Akbas, "Self-supervised learning of 3D human pose using multi-view geometry," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1077–1086.
[63] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy, "A multi-view RGB-D approach for human pose estimation in operating rooms," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 363–372.
[64] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[65] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.
[66] A. Doumanoglou, P. Drakoulis, N. Zioulis, D. Zarpalas, and P. Daras, "Benchmarking open-source static 3D mesh codecs for immersive media interactive live streaming," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 190–203, Feb. 2019.
[67] *Corto*. Accessed: May 17, 2020. [Online]. Available: https://github.com/cnr-isti-vclab/corto#libcorto
[68] Google Inc. *Google Draco*. Accessed: May 17, 2020. [Online]. Available: https://github.com/google/draco
[69] Q. Meyer, J. Süßmuth, G. Sußner, M. Stamminger, and G. Greiner, "On floating-point normal vectors," in *Computer Graphics Forum*, vol. 29, no. 4. Hoboken, NJ, USA: Wiley, 2010, pp. 1405–1409.
[70] P. Cignoni, C. Rocchini, and R. Scopigno, "METRO: Measuring error on simplified surfaces," *Comput. Graph. Forum*, vol. 17, no. 2, pp. 167–174, Sep. 1998.
[71] R. B. Rusu and S. Cousins, "3D is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 1–4.
[72] S. Subramanyam, J. Li, I. Viola, and P. Cesar, "Comparing the quality of highly realistic digital humans in 3DoF and 6DoF: A volumetric video case study," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Mar. 2020, pp. 127–136.
[73] M. 3DG and Requirements, *Call for Proposals for Point Cloud Compression*, Standard ISO/IEC JTC1/SC29 WG11 N16732, Geneva, Switzerland, Jan. 2017.
[74] S. Schwarz, M. Preda, V. Baroncini, M. Budagavi, P. Cesar, P. A. Chou, R. A. Cohen, M. Krivokuca, S. Lasserre, Z. Li, J. Llach, K. Mammou, R. Mekuria, O. Nakagami, E. Siahaan, A. Tabatabai, A. M. Tourapis, and V. Zakharchenko, "Emerging MPEG standards for point cloud compression," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 133–148, Mar. 2019.
[75] D. Tian, H. Ochimizu, C. Feng, R. Cohen, and A. Vetro, *Updates and Integration of Evaluation Metric Software for PCC*, document MPEG2017 M, ISO/IEC JTC1/SC29/WG11, 2017, vol. 40522.
[76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

ANARGYROS CHATZITOFIS (Graduate Student Member, IEEE) received the Diploma degree in electrical and computer engineering (ECE) from the ECE Department, National Technical University of Athens (NTUA). His main research expertise lies on human-centric 3D vision and machine learning. His Ph.D. research was focused on depth-based motion capture and deep learning at NTUA ECE. He has (co)authored more than 20 scientific publications in international computer vision conferences and journals.

LEONIDAS SAROGLOU graduated from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (A.U.Th.). Since then, he has been working as a Research Assistant with the Information Technologies Institute (ITI), Centre for Research & Technology Hellas (CERTH). His research interests include image processing, pattern recognition, real-time 3D reconstruction, 3D computer vision, and deep learning.

PRODROMOS BOUTIS graduated from the Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki (A.U.Th.), in July 2018. Since February 2019, he has been working as a Research Assistant with the Information Technologies Institute (ITI), Centre for Technological Research & Technology Hellas (CERTH). His research interests include 3D computer vision, digital image processing, 3D model rendering, pattern recognition, and machine learning.

PETROS DRAKOULIS received the B.Sc. degree in IT engineering from Alexander TEI, and the M.Sc. degree (Hons.) in digital media and computational intelligence from the Aristotle University of Thessaloniki. In 2018, he joined the Visual Computing Laboratory, ITI-CERTH, where he works as a Research Assistant and Software Developer. His research interests include software engineering, visual computing, machine learning, and graphics.

NIKOLAOS ZIOULIS received the degree from the Aristotle University of Thessaloniki, in 2012. He has been working as an Electrical and Computer Engineer with the Information Technologies Institute (ITI), Centre for Research & Technology Hellas (CERTH), since October 2013. His research interests include computer vision and graphics technologies, and more specifically in volumetric 3D capturing and rendering, 3D scene understanding, and tele-immersive applications.

SHISHIR SUBRAMANYAM (Graduate Student Member, IEEE) received the B.Tech. degree in computer science from BITS Pilani Dubai, and the M.Sc. degree in computer science from the Delft University of Technology. He is currently pursuing the Ph.D. degree with the Distributed and Interactive Systems Group, Centrum Wiskunde & Informatica. His research interest includes multimedia systems, specifically on the transport and delivery of volumetric media.

BART KEVELHAM received the M.Sc. degree in computer science from the University of Twente, The Netherlands, in 2006, specializing in computer graphics. He is currently a Lead Research and Development Engineer with Artanim, Geneva, Switzerland, where his work focuses on the research and development of solutions enabling interactive full-body and free-roam VR experiences. His research interests include real-time computer graphics, physical simulation, and computer vision.

CAECILIA CHARBONNIER received the Ph.D. degree in computer science from the MIRALab, University of Geneva, Switzerland, in 2010. She is the Co-Founder and the Research Director of Artanim, a center specialized in motion capture technologies, and the Co-Founder and CIO of Dreamscape Immersive, a VR entertainment company. Her research interests include the interdisciplinary use of motion capture from 3D animation, live performances to movement science, orthopedics, and sports medicine.

PABLO CESAR (Senior Member, IEEE) currently leads the Distributed & Interactive Systems Group, Centrum Wiskunde & Informatica (CWI). He is also an Associate Professor with TU Delft, The Netherlands. His research combines HCI and multimedia systems, and focuses on modeling and controlling complex collections of media objects distributed in time and space. He is a member of the Editorial Board of *IEEE Multimedia*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and the IEEE TRANSACTIONS ON MULTIMEDIA, among others.

DIMITRIOS ZARPALAS received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki (A.U.Th.), the M.Sc. degree in computer vision from Pennsylvania State University, and the Ph.D. degree in medical informatics from the Department of Medicine, Health Science School, A.U.Th. In 2008, he joined the Information Technologies Institute (ITI), Centre for Research & Technology Hellas (CERTH), as an Associate Researcher, where he is currently a Senior Researcher (Grade C).

STEFANOS KOLLIAS (Fellow, IEEE) has been a Professor with the ECE School, National Technical University of Athens, since 1997. He has also been a Professor of Machine Learning with the Computer Science School, University of Lincoln, U.K., since 2016. He has published 110 journal articles and 310 conference papers. He has supervised 43 Ph.D. students. His research interests include machine and deep learning, multimedia analysis, search, retrieval and recognition, vision, medical informatics, cultural heritage, HCI, and affective computing.

PETROS DARAS (Senior Member, IEEE) received the Diploma degree in electrical and computer engineering and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is currently the Research Director and the Chair of the Visual Computing Lab, coordinating the research effort of more than 80 scientists and engineers. His main research interests include visual content processing, multimedia indexing, and machine learning. His involvement with those research areas has led to the coauthoring of more than 300 papers in refereed journals and international conferences.