SmartPortraits: Depth Powered Handheld Smartphone Dataset of Human Portraits for State Estimation, Reconstruction and Synthesis

Anastasiia Kornilova Marsel Faizullin Konstantin Pakulev Andrey Sadkov
Denis Kukushkin Azat Akhmetyanov Timur Akhtyamov Hekmat Taherinejad
Gonzalo Ferrer
Center for AI Technology (CAIT), Skolkovo Institute of Science and Technology

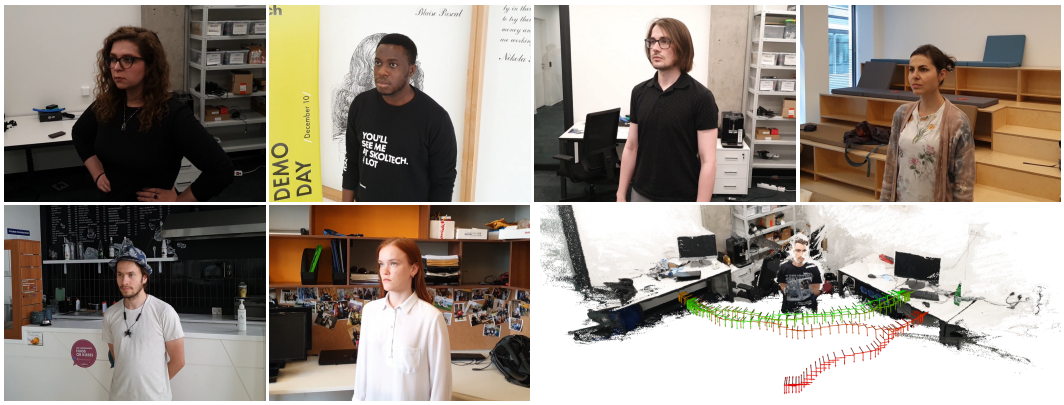

Figure 1. From *top-left:* examples of frames from the SmartPortraits dataset videos that capture human portraits in different natural environments, with varying lightning conditions, using a smartphone and external depth camera on a rig. *Bottom-right:* recorded trajectory (red – initial time, green – end time) and dense reconstruction obtained by ACMP [88].

Abstract

We present a dataset of 1000 video sequences of human portraits recorded in real and uncontrolled conditions by using a handheld smartphone accompanied by an external high-quality depth camera. The collected dataset contains 200 people captured in different poses and locations and its main purpose is to bridge the gap between raw measurements obtained from a smartphone and downstream applications, such as state estimation, 3D reconstruction, view synthesis, etc. The sensors employed in data collection are the smartphone's camera and Inertial Measurement Unit (IMU), and an external Azure Kinect DK depth camera software synchronized with sub-millisecond precision to the smartphone system. During the recording, the smartphone flash is used to provide a periodic secondary source of lightning. Accurate mask of the foremost person is provided as well as its impact on the camera alignment accuracy.

For evaluation purposes, we compare multiple state-of-the-art camera alignment methods by using a Motion Capture system. We provide a smartphone visual-inertial benchmark for portrait capturing, where we report results for multiple methods and motivate further use of the provided trajectories, available in the dataset, in view synthesis and 3D reconstruction tasks.

1. Introduction

Realistic rendering of people, and in general of objects, has recently achieved an unprecedented level of detail and realism [4, 27, 46, 49, 76, 87, 93, 94] with potential groundbreaking applications in telepresence, VR and AR. Most of these methods have focused on static scenes and synthetic data, leaving apart the computational time required, which is still prohibitive. In contrast, many potential usages of reconstruction and rendering are ideal candidate applications for smartphones, or other consumer-level devices, whose sensors are improving every year but still of limited quality. Our objective is to create a dataset that recreates *in the wild* conditions emulating smartphone users.

The SmartPortrait dataset[1] is an effort to bridge the gap between realistic raw data obtained from people, collected

[1] https://MobileRoboticsSkoltech.github.io/SmartPortraits/

from a handheld smartphone and the down-stream reconstruction applications, for instance 3D portrait reconstruction, view synthesis, etc. The key component that links both views is camera pose *state estimation*. A usual practice is to obtain these poses by using a reliable but computationally demanding Structure from Motion (SfM) algorithm such as COLMAP [71] or multi-modal SLAM methods [13, 41, 68, 80]. The trajectories provided in our dataset emulate handheld movements, as if the user, or close-by users, were recording the sequences (see Fig. 1).

Many view synthesis methods [28, 49, 57, 74] generate their own datasets just by using state estimation methods. These single camera free-viewpoint images can only be considered if the scene is *static*. We asked volunteers to stay as still as they could while recording them in a semicircular trajectory from close and mid distances. We observed that most of the volunteers slightly changed their postures so we should expect some degree of displacement, which transforms the problem into *non-static*.

The SmartPortrait dataset is obtained in a variety of emplacements, under different lightning conditions, plus a flashing light from the smartphone at regular intervals. The smartphone camera is complemented with a high-quality depth sensor, adding robustness and multi-modality.

We provide a recorded dataset consisting of smartphone video images, IMU data, perfectly time aligned, and an external depth camera from Azure Kinect DK. The evaluation includes two steps: first we compare the most promising methods with a reference trajectory obtained from a motion capture (MoCap) system. Second, for some environments, it is not possible to deploy the MoCap system. Therefore, we provide a reference trajectory, obtained from the previous best performing method, and provide an upper bound of the error by using a non-reference metric [40]. In further evaluations, we benchmark multiple state of the art methods for visual SLAM, SfM and Visual-Inertial based methods.

Next, we want to connect the problem of camera pose estimation with two downstream tasks: 3D reconstruction with COLMAP [72], ACMP [89] and SOTA view synthesis algorithms (NeRF [49], FVS [66], SVS [67]). These applications will help us to understand the importance of pose estimation and its correlation with other tasks.

Ethical considerations. We asked all participants in the dataset for a signed consent to record their portraits and publicly release them for purely academic purposes. We explicitly indicated in the agreement their right, at any time, of removing all their data.

2. Related work

Capturing human data is always addressed to a particular task, for instance, human faces, portraits, facial expressions, hand gestures, full bodies, etc. The common trait is that obtaining these data is a challenging process and different approaches exists to tackle them. Our data include human portraits, or upper body of people, and there exists a relation to many other works on capturing human data. This section reviews the existing literature based on the sensor set used to obtain them. Later, we will discuss some applications and finally, we will present some state estimation methods as a requirement for single free-viewpoint recordings.

Motion Capture systems [1, 3] utilize multiple customized cameras to accurately detect reflective or infra-red markers. They are a popular method used to capture human data by tracking markers and synchronize them with video: HumanEva [77], Human3.6M [35] and INRIA [90]. A negative effect is that it requires the volunteers to wear special suites over their bodies, changing their clothing appearance.

Multiple cameras overcome this issue and remove the need for *markers*. They are a very popular method to capture body expressions and fine details, preserving visual appearance of the models. Examples include shape capture [85], streamable free-viewpoint [17], AIST [83], Panoptic studio [38, 39] with a mixture of 500 cameras, BUFF [95] for human pose and shape estimation, Humbi [92] for body expressions, [28, 87] for head portraits, or photo-realistic full-body avatars [8]. These settings are in practice very precise at capturing simultaneously the same event, e.g. a dynamic person. They are however expensive, difficult to deploy in different environments, and require a considerable amount of effort to calibrate and synchronize them.

Controlled **lightning** conditions are also becoming an important feature for obtaining a fine detailed geometry reconstruction when digitizing humans [32, 74, 85]. SmartPortraits includes some images under smartphone flash conditions, such that the lightning source coincides with the optical sensor frame and creates a different outcome than under ambient lightning.

Some attempts have tried to lower the demanding requirements of multi-camera settings, where many sensors and lightning sources are required. One solution is enhancing the data obtained with a single **depth sensor** [10, 33, 42, 75, 91] or multiple depth sensors [21, 34, 96]. Other approaches try to reduce the number of cameras in operation and still obtain reasonably accurate results [98, 99].

At the extreme, one would desire a **single camera** free-viewpoint, either taking multiple pictures or with a video [5, 89]. This is the aim of our dataset as well.

From the perspective of datasets capturing human data for people reconstruction and rendering task, we observe the following modelling classes: full body modelling (DynamicFAUST [11], BUFF [95], People Snapshot [6]), clothes modelling (3DPeople [59], SIZER [81]), *head/torso portrait* modelling (UHDB11 [82], Nerfies [57], Portrait Neural Radiance [57]), or suitable for applications in couple of tasks (RenderPeople, Humbi [92]). There are also crowd-sourced datasets such as MannequinChallenge [44], TikTok

	EuRoC MAV	TUM-VI	TUM RGB-D	PennCOSYVIO	KAIST VIO	ADVIO	Ours
Year	2016	2018	2012	2016	2021	2018	2021
Environment	indoors	indoors/ outdoors	indoors	indoors/ outdoors	indoors	indoors/ outdoors	indoors
Carrier	MAV	handheld	handheld/robot	handheld	UAV	handheld	handheld
Focus	MAV VIO/SLAM	VIO	RGB-D SLAM	handheld VIO	UAV VIO	handheld VIO/SLAM	VIO/SLAM in Human Digitiz.
Cameras	stereo gray: 2x752x480 @20Hz	stereo gray: 2x1024x1024 @20Hz	RGB-D: 640x480 @30Hz	• 4 RGB: 1920x1080 @30Hz • stereo gray: 2x752x480 @20Hz • fisheye gray: 640x480 @30Hz	• RGB: 640x480 @30Hz • stereo IR: 640x480 @30Hz	• RGB: 1280×720 @60Hz • fisheye gray: 640x480 @60Hz	• RGB: 1920x1080 @30Hz • depth: 640x576 @5Hz
IMUs	ADIS16448 3-axis acc/gyro @200Hz	BMI160 3-axis acc/gyro @200Hz	Kinect 3-axis acc @500Hz	• ADIS16488 3-axis acc/gyro @200Hz • Tango 2 3-axis acc @128Hz • Tango 2 3-axis gyro @100Hz	Pixhawk 4 Mini 3-axis acc/gyro @100Hz	MP67B 3-axis acc/gyro @100Hz	• LSM6DSO 3-axis acc/gyro @500Hz • MPU-9150 3-axis acc/gyro @500Hz
Time sync	hw	hw	hw	hw, sw	data	sw	hw, sw + frame sync
Point clouds	✓(some seq)	×	✓	×	×	✓	✓
Distance	11 seq, 0.9 km	28 seq, 20 km	39 seq x several m	4 seq, 0.6 km	14 seq x several m	23 seq, 4.5 km	1000 seq, 6.6km
Ground-truth	• 3D pos. (some seq), laser tracker @20Hz • 3D pose (some seq), MoCap @100Hz • 3D pcds (some seq), laser tracker	3D pose, MoCap @120Hz (partial gt)	• 3D pose, MoCap @300Hz (partial gt) • 3D pcds, Kinect @5Hz	3D pose, visual markers @30Hz	3D pose, MoCap @50Hz	• 3D pose, IMU + manual position fixes @100Hz • 3D pcds, Tango @5Hz	• 3D pose (some seq), MoCap @240Hz • 3D pose, COLMAP/RGB-D SLAM @5Hz
Acc. ≈	1 mm	1 mm (static case)	1 mm (relative)	15 cm	1 mm	0.1 - 1 m [86]	1 mm - 1 cm

Table 1. Overview of common Visual (V) and Visual Inertial (VI) benchmark datasets targeted at state estimation.

Dataset [36] collected from social networks available for reconstruction tasks. Our dataset is unique since it records consumer-level data (smartphone) "in the wild" supported by high-quality external depth data.

Recently emerged neural implicit representation methods allow to bypass the need for obtaining accurate 3D structure of a scene and instead model it implicitly, e.g. by considering occupancy [48], signed distance function [56] or volumetric density [46, 49]. In particular, there have been several works that successfully used neural implicit representation for creating realistic portrait avatars [27, 28, 57, 74, 87].

Unfortunately, when scenes include dynamic elements, and that is the case of video recording of people, the state estimation of camera poses or free-viewpoints and the 3D scene is not so trivial. In the downstream tasks of 3D human reconstruction, two main variants exist: free-form [15, 69, 70] and model based [6, 47, 55].

Accordingly, when reducing the number of cameras to a single one and working in *non-static* conditions because the volunteers in our dataset stand still but not immobile, then, **state estimation** becomes the key ingredient that allows a handheld single camera video to be used for human digitization. Either if camera poses are compensated while learning a model [45] or estimated as recorded, the quality of these is going to be determinant for any downstream task.

To the best of our knowledge, there are no datasets that directly address the evaluation of state estimation approaches when a human is in the main focus of sensors. State estimation of camera poses include techniques such as Visual Odometry (VO) [22, 25], Visual-Inertial Odometry (VIO) [9, 41, 43, 62]. Variants of Simultaneous Localization and Mapping (SLAM), which include a loop in estimation for global pose alignment, either Visual SLAM (V-SLAM) [29, 51, 52, 80] or Visual-Inertial SLAM (VI-SLAM) [13, 31, 61, 68] and Structure from Motion (SfM) [71], all of them relevant to the be applied to the sensor data available in a smartphone. To date, there are numerous datasets available [12, 14, 16, 18, 19, 30, 37, 58, 73, 79, 86, 100] that vary greatly by their focus, recording environment, sensor carriers as well as by the amount of data recorded and the accuracy of the ground truth. We briefly describe the main features of the main datasets and give a comparison with our dataset (see Table 1).

The EuRoC Micro Aerial Vehicles (MAV) dataset [12] focuses on VIO and SLAM for MAVs as well as 3D reconstruction. The authors employ a stereo pair of cameras hardware-synchronized with an IMU installed on a MAV for acquiring the data sequences in two indoor environments. Kaist VIO [37] is another indoor dataset that focuses on VIO for aerial vehicles, it specifically addresses challenging scenarios for VIO that contain pure rotational/harsh motions. TUM-VI [73] is a dataset for the purpose of evaluating VIO algorithms. Compared to other mentioned datasets it stands out by its size, diversity of the recorded sequences, as well as uses higher resolution cameras. TUM RGB-D [79] features only indoor sequences captured with a Kinect sensor that is handheld or mounted on a robotic platform. The dataset includes challenging scenarios for the proper evaluation of RGB-D SLAM approaches. PennCOSYVIO [58] is one more VIO benchmark that contains diverse challenging sequences. It includes not only rotational motions but hard visual conditions as well. The dataset was captured using a larger number of sensors than any other related datasets: 3 GoPro cameras, an integrated VI-sensor and 2 Google Project Tango tablets mounted on a

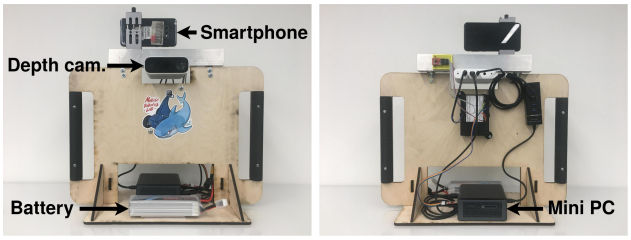

Figure 2. Front and back view of the recording platform.

rig. However, as pointed out in [73] the setup yields lower synchronization accuracy between cameras and IMUs when compared to datasets like TUM-VI. ADVIO [18] focuses on benchmark VIO and SLAM methods for smartphones and mobile devices with low-cost sensors. It contains different large indoor and outdoor environments recorded in public places.

Our dataset contains recordings of people in indoor environments and focuses on the diversity of people, their clothing, the environments and the lightning conditions, totally uncontrolled aiming to recreate every-day life conditions of smartphone users. Compared to PennCOSYVIO, KAIST VIO and ADVIO that use heterogeneous sensors too, we employ a significantly more precise hybrid hw/sw synchronization technique from [24] (see Sec. 3.1). Due to recordings surroundings peculiarities we provide pseudo-ground truth poses based on the gathered sensors data like in [18] the quality of whose is estimated in a manner similar to [58] by capturing verification sequences (see Sec. 5.2).

3. Recording Platform

Our dataset aims to provide human portrait data in the realistic environments captured by a middle-price smartphone. To meet these requirements, we have designed a portable handheld platform with a **Samsung S10e** smartphone (RGB camera, 1920x1080 p, 30 fps; IMU, 500 Hz, flasher, 1 Hz) and a high-end depth sensor **Azure Kinect DK** (depth camera, 640x576 p, 5 fps). The high-quality external depth camera is chosen instead of the smartphone with a built-in sensor as (i) the modern smartphone depth images are still not as high-quality as external depth sensors, and (ii) it is not possible to record RGB and depth tracks by the smartphone on high frequency simultaneously. A common view of the system is presented in Fig. 2.

Specifics of our recording case — dynamic camera movements close to real-life handheld capture and a person in foreground with non-stationary pose (blinking eyes, small movements of the person because of breath, coordination, heart beats).

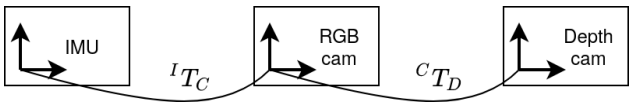

Figure 3. Obtained relative transformations by calibration. $^{I}T_{C}$ – smartphone camera in smartphone IMU reference frame, $^{C}T_{D}$ – depth camera in smartphone camera frame. $^{C}T_{D}$ is found as $^{C}T_{D} = ^{C}T_{R} \cdot ^{R}T_{D}$, where $^{C}T_{R}$ is Azure RGB camera in smartphone camera frame obtained by Kalibr, $^{R}T_{D}$ is factory-known Azure RGB camera in Azure depth camera frame. Azure RGB is only used for this procedure.

3.1. Time and Frame Synchronization

The independence of smartphone and depth camera adds an additional challenge to the time synchronization. If frames from both sensors are captured at slightly different instants of time, several tens of ms, this degrades the quality of the camera pose estimation.

To synchronize the cameras, we introduce a two-step sync process. First, the time domains between two sensors are synchronized by Twist-n-Sync algorithm [23] which is not affected by network asymmetry, in contrast to network-based protocols like NTP. And in the second step – the synchronization of frame capturing moments from both sensors is done. For solving that, the grabbing of the smartphone's framing phase is performed via remote API interface. Then depth camera triggering is automatically tuned to this phase as explained in [24]. The sync used provides sub-millisecond accuracy.

3.2. Calibration

The full intrinsic and extrinsic calibration of smartphone camera and smartphone IMU, is obtained by the Kalibr toolbox [65] with 6x6 AprilGrid array of 3x3 cm AprilTags [54] as the visual markers.

To find depth-to-smartphone camera transformation, firstly, we obtained the Azure RGB to smartphone camera transform. Then, combined it with the factory-known Azure depth (infra-red) to Azure RGB camera transform. This method gives much better accuracy than direct depth-to-smartphone camera transformation with low-quality infrared camera. We use Azure RGB only for this procedure. All the obtained transformations are shown in Fig.3.

Smartphone camera is rolling-shutter type; however, we applied global-shutter camera model during calibration to feed the calibration parameters correctly to the methods we compared in Sec. 5. Standalone IMU calibration is also performed. IMUs noise parameters were borrowed from [78] and [53] for smartphone and standalone IMU respectively.

4. Dataset

Our dataset contains 1000 records of 200 people, with natural clothing, captured in different native locations and poses. Every record consists of synchronized smartphone data (Full HD RGB video, flash timestamps, timestamped accelerometer and gyroscope measurements) and the depth images from external high-quality depth sensor. The dataset also contains reference ground truth trajectories of the smartphone camera obtained as described in Sec. 5.2 and segmentation masks for the person. We supplement the dataset by labels for genders and difficult cases for rendering like volume hairs/beard/glasses. Data parameters and sampling rates of the sensors are presented in Table. 1.

4.1. Collection Process and Statistics

During every recording process, three people are involved: (1) a volunteer who is being filmed, (2) an operator that carries the recording platform, and (3) an assistant that monitors the correctness of the recording through SSH. The volunteer is asked to stand or sit still. The operator carries a recording platform around the person at the subject's face height as depicted in Fig. 1. The recording trajectory begins in front of the person to capture the whole scene, then the operator moves to a side of the volunteer and makes four 100-120 degrees circular arcs around the model. The entire trajectory is shown in Fig. 1. The timestamps of every arc edge are marked online during recording by the assistant in an automated manner. In post-processing stage, the whole trajectory could be split into separate arcs applying these marks.

Every person is captured in 5 different poses — 3 in a standing position (straight, hand on hips, with head turned) and 2 in a sitting position (straight, with head turned). Standing and sitting positions were captured from a distance of about 2 and 1 m respectively. During the recording, blinking flash on the smartphone is turned on with a frequency of 1 Hz to relight the model. Effect of re-lightning is more distinguishable on sitting positions since they are captured from a closer distance.

Data collection is performed in 5 different locations of native indoor environments: cafeteria, lab, office, campus entrance, and student council. Their common view is demonstrated in Fig. 1. The average length of the trajectories for staying and sitting position are 7.14 and 5.8 m accordingly. The total duration of all tracks is 11 hours and 6 minutes, and the total length is 6610 meters.

SmartPortrait contains people of different gender, appearance, clothing, hairstyles, etc. Statistics are shown in Fig. 4.

4.2. Segmentation Masks

Along with recorded data, we also provide segmentation masks of humans on the images. This information could be used for filtering out potentially dynamic landmarks of the scene on the trajectory estimation step (blinking eyes, subject movements) as demonstrated in Sec. 5.2 or for separation of portraits parts from the scene for only-person 3D reconstruction. For this task, we design a semi-automated labeling process, based on U2-Net [63] that is pre-trained on people masks from the Supervisely Person Dataset [2]. Usage of this method on our data overestimates the person mask, also covering some parts of the background. DBSCAN clusters the masked part by using the depth component, discarding scene parts that are not related to the foreground. Finally, segmentation results are assessed visually by labelers.

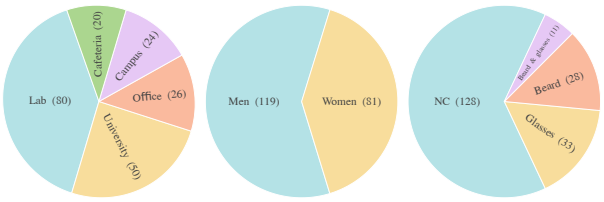

Figure 4. Dataset statistics. *Left:* Locations of recording. *Center:* Gender *Right:* Appearance.

5. Evaluation

The evaluation part tackles two main questions — (1) how to find the best way of calculating pseudo-ground truth poses for our dataset and (2) investigate the performance of V and VI state estimation methods on smartphone data only. V denotes all visual methods: VO and V-SLAM; the same applies for VI.

5.1. Metrics

Full-reference metrics. Among the class of full-reference metrics where the reference trajectory (ground truth) is available, we consider RMSE statistics on Absolute Pose Error (APE) and Relative Pose Error (RPE) for the rotation and translation parts. In particular, for translation APE, we apply the Umeyama alignment [7, 84] between a pair of trajectories if expressed in different origin frames. For rotation APE, the Umeyama alignment is followed by the trajectory's reference frame transformation.

No-reference metrics. No-reference metrics are alternative to the full-reference metrics when the reference trajectory is not available or its quality is disputable. In our work, we use Mutually Orthogonal Metric (MOM) [40] that measures quality of the trajectory by evaluating quality of the map aggregated from point clouds registered via the trajectory poses. MOM provides stronger correlation with RPE error in comparison to its competitors [64]. In our setting, MOM uses the point clouds converted from depth images.

In order to apply MOM on trajectories ambiguous to

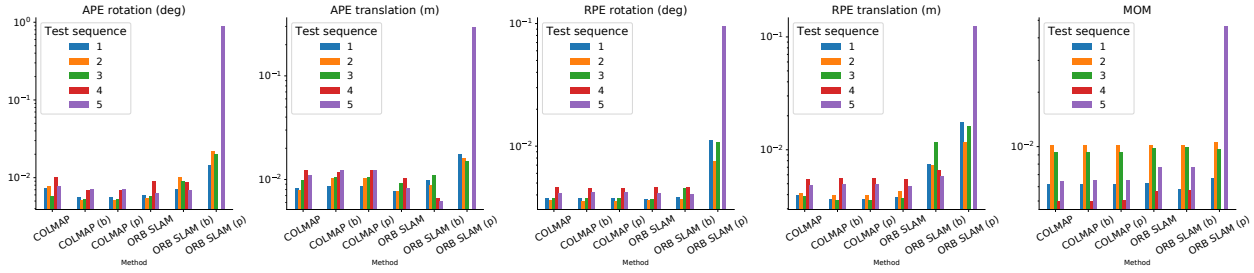

Figure 5. Full-reference (APE/RPE) and no-reference (MOM) metric statistics for COLMAP and ORB SLAM (RGB-D) for 5 test sequences with MoCap ground truth poses. (b) and (p) indicate that only background keypoints and person keypoints correspondingly were considered for pose estimation. One pose for ORB SLAM (p) is not converged, therefore its values are excluded from the evaluation.

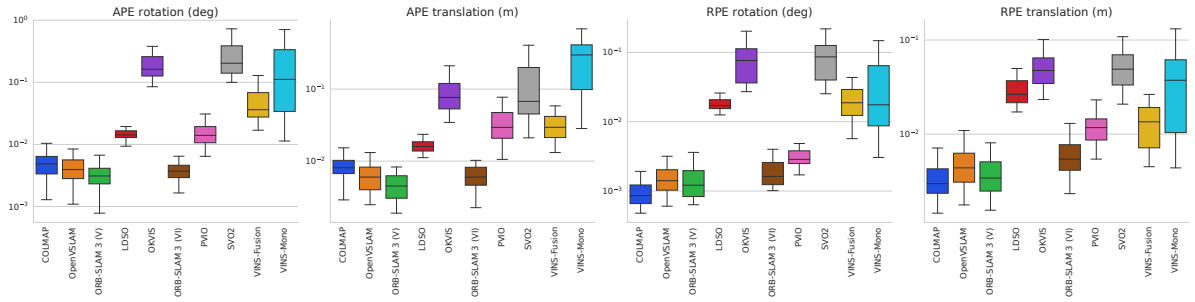

Figure 6. Evaluation of V/VI methods that employ only smartphone data (frames and IMU) on benchmark sequences.

scale (e.g., COLMAP), we optimize the scale factor w.r.t. MOM metric — which assumes that the correct value of scale is reached at the optimum in the metric, when the aggregated map of point clouds is at its best condition.

5.2. Ground Truth Trajectories

The majority of the dataset sequences (see Fig. 4) are captured in public places or areas where either applying conventional methods of acquiring ground-truth poses (e.g., MoCap) are not feasible, or such methods disrupt the nativity of the surroundings (e.g., visual markers).

Therefore, it is required a no-reference method in order to validate the obtained trajectories when MoCap data is not available. Below, we will present a procedure to select a new reference trajectory and an upper bound of its error.

Methods. Since the dataset is targeted on both state estimation and reconstruction/synthesis domains, we consider the main methods typically used by the community that make use of the sensors: RGB, depth cameras and IMU. From *reconstruction and rendering* field experience, we consider the COLMAP [71] Structure-from-Motion (SfM) pipeline that is de-facto standard tool in this area and usually employed as ground truth. COLMAP uses only RGB data and therefore its trajectory is defined up to a scale factor, that limits its usage in state estimation tasks. In addition, we consider the class of RGB-D SLAM algorithms that are able to provide poses and scale is observable. Based on the wide evaluation of RGB-D SLAM methods done in [97] we choose ORB-SLAM (RGB-D) [52], implementation from [13], as one with the lowest trajectory error.

MoCap Test Sequences. To assess the accuracy of the ground truth poses, we record several testing sequences in the laboratory environment where the use of a more accurate ground truth acquisition method is possible. In particular, we utilize OptiTrack MoCap system [1] to record 5 testing sequences of one person in the common dataset format. MoCap is synchronized with the platform offline by the Twist-n-Sync algorithm [23]. The extrinsic parameters calibration requires to calculate:

$$\min_{X,Y \in SE(3)} \sum_i || \log \left(Y \cdot T_M(i) \cdot X \cdot T_C^{-1}(i) \right) ||, \qquad (1)$$

where $T_M(i)$ is the trajectory given by the MoCap at time i, T_C is pose at the camera frame calculated by the algorithm, X is the transformation between the camera optical center and the tracked object in the MoCap and Y is the transformation between the origin frame of the MoCap and the origin of the SLAM algorithm (usually first frame).

Results. In order to support the selection of the pseudo ground truth, we evaluate COLMAP and ORB-SLAM (RGB-D) (actually, virtual stereo) on the MoCap test sequences by using the described above full-reference and no-reference metrics. Because the landmarks on the volunteer body might be non-static (person can breath, blink),

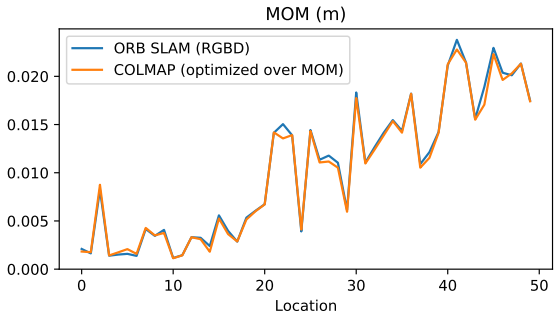

Figure 7. MOM generalization to 50 other scenes from the dataset.

we consider three modifications of COLMAP and ORB-SLAM (RGB-D) – using the whole scene, using mask for background, using mask for person. The evaluation results are presented in the Fig. 5. Both methods demonstrate almost the same performance on the considered metrics, being close to the resolution limit of the MoCap system. The usage of masks does not affect the performance in case of COLMAP that could be explained by photometric consistency imposed by the algorithm. ORB SLAM performs worse when only part of the scene is visible.

Eval Generalization. The above evaluation on the MoCap test sequences with ground truth available is limited only to the lab location. To extend it to all locations that our dataset covers, we must consider the comparison of COLMAP and ORB-SLAM (RGB-D) by using the no-reference metric MOM. For that, we select 10 trajectories from every location that result into 50 test sequences. Evaluation performance is presented in Fig. 7 and it allows to make the following two conclusions. Firstly, since MOM measures the dispersion of deviations on planar surfaces in the aggregated map it can be noticed that for the majority of locations it does not overcome 2 cm. This value is comparable to the depth sensor noise, that means that both methods give relatively good trajectories from the state estimation point of view. Secondly, COLMAP performs slightly better than ORB-SLAM (RGB-D), although it requires post-processing for revealing the scale. We will provide the obtained trajectories of both methods as the pseudo-ground-truth methods.

5.3. V and VI Evaluation

One of the motivations of our work is the study the potential of applications of V/VI methods using smartphone-only data targeted to the domain of human portraits. In this section, we provide evaluation of different state-of-the-art methods and a baseline for future comparisons. In addition, to all the considered methods, we deliver configuration and calibration files for methods to be used with our data for benchmark.

Methods. In evaluation we consider two classes of methods: Visual (V) and Visual Inertial (VI) methods. Considering recent exhaustive evaluations [20, 37], we order top-rated V/VI methods. The considered methods for both classes are: for V — OpenVSLAM [80], ORB SLAM Monocular [13], LDSO [29] and COLMAP [71]. For VI (ordered by performance on other datasets) – ORB-SLAM 3 (VI) [13], Kimera VIO [68], OpenVINS [31], VINS-Fusion [62] [60], VINS-Mono [61], PVIO [43], SVO.2 [26], MSCKF [50], OKVIS [41], ROVIO [9]. Some of the methods (Kimera VIO, OpenVINS, MSCKF, ROVIO) were discarded because they require recording device to be static over the first seconds of the trajectory for initialization, whereas our use case does not cover such scenario.

Benchmark Dataset To evaluate the performance of the methods, we uniformly pick 2 sequences for every combination of location and volunteer pose, resulting in a total of 50 evaluation sequences. As ground truth, we consider the trajectories produced by ORB-SLAM (RGB-D) that, as demonstrated in Sec. 5.2, provides excellent performance.

Results. The evaluation results on the set of the full-reference metrics are presented in Fig. 6. Because the pseudo GT provides a statistical bound, it could be used for exact ordering in cases when the order of error magnitude is higher than the error between MoCap and pseudo GT. In particular, in our comparison we can order only the next V/VI methods: LDSO, OKVIS, PVIO, SVO2, VINS. In general, we can observe that V methods perform more accurately in rotation and translation than VIO methods. The VIO method's accuracy varies, where ORB 3 VI performing as accurate as V methods. All VI have better than 1 degree of accuracy in absolute rotation error.

6. Application

3D reconstruction. For qualitative comparison on 3D reconstruction, we provide poses obtained from the top performing methods for state estimation from every class and use two state-of-the-art methods for 3D reconstruction: COLMAP multi-view stereo (MVS) [72] and ACMP [88]. The demonstration of the obtained 3D scenes is presented in Fig. 8 from two views — center and the edge of the trajectory arcs. For both reconstruction pipelines, COLMAP trajectory produces less distorted reconstruction, overcoming ORB-SLAM (RGB-D) and (V) versions. It could also be noticed that from the trajectory border point of view 3D reconstruction has less quality. The solution from the VI method produced an incorrect reconstruction, perhaps due to its error.

View Synthesis. We provide a qualitative comparison of the considered VO-SLAM methods evaluated on an image synthesis problem. For that we considered SOTA methods: Neural Radiance Field [49] (NeRF), and generalized NeRF approaches — FVS [66] and SVS [67] in their de-

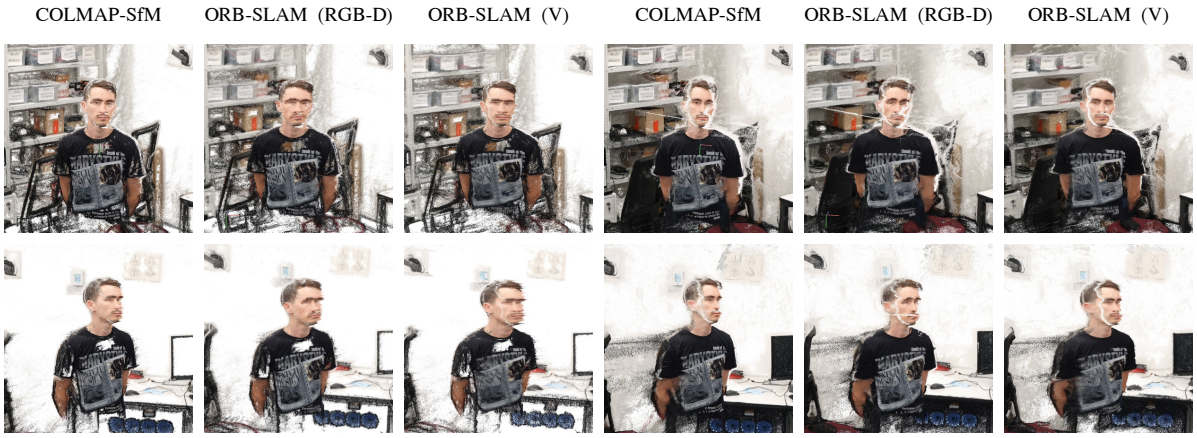

Figure 8. Qualitative demonstration of dense reconstruction from different views using COLMAP-MVS (3 left columns) and ACMP (3 right columns) on poses from COLMAP-SfM, ORB-SLAM (RGB-D), ORB-SLAM (V).

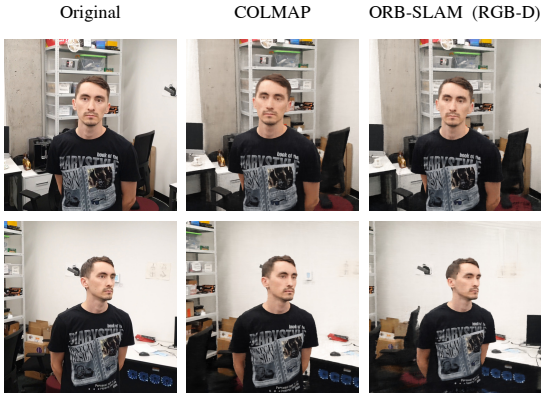

Figure 9. Qualitative demonstration of the NeRF novel-view synthesis algorithm on new poses, not observed before.

Figure 10. Qualitative demonstration of the FVS and SVS novel-view synthesis algorithms on new poses, not observed before.

fault pre-trained versions. The data provided to methods is the solution of the trajectories. For NeRF algorithm, that is optimized per each scene, COLMAP provides the best qualitative results, the objects are coherently synthesized, a little blurred. ORB-SLAM (RGB-D) result shows some inconsistencies on the rendering (Fig. 9-*Bottom-right*) and overall less definition than COLMAP, although they both showed a similar trajectory error. Results of FVS and SVS are presented in Fig. 10. For both, COLMAP and ORB, their provide less quality than NeRF. It could be explained by difference in poses configuration w.r.t original methods data on which methods were trained.

7. Discussion

One of the questions that this paper raised in the introduction is: *Are we ready to calculate handled trajectories in the wild and convert them into 3D portraits of people?*

Real time VI methods do not perform as accurately as V methods (some of them not real-time). Still, the accuracy achieved is remarkable, providing very accurate trajectories; probably, they could be better if IMUs were properly initialized.

Despite their accuracy, the results obtained on the applications (NeRF and reconstruction) could be improved. This question is to be explored, but our qualitative results hint that trajectory error is not perfectly correlated with the downstream task, either for synthesis or for reconstruction. An explanation can be the photo-metric consistency, which is more important than the trajectory error. A corollary of this: perhaps the solution is not to jointly optimize trajectories and maps, but simply optimize the map, allowing for small disturbances on the camera poses from a reasonable initial solution. We plan to further investigate this issue.

Potential Negative Societal Impact. Realistic human data are required to achieve a future of immersive VR and tele-presence. However, there are other potentially dangerous uses of these technology such as identity theft or fake news.

Acknowledgments. This research is based on the work supported by Samsung Research, Samsung Electronics.

References

[1] OptiTrack. www.optitrack.com/. 2, 6

[2] Supervisely person dataset. https://supervise.ly/explore/projects/supervisely-person-dataset-23304/datasets. 5

[3] Vicon motion systems. www.vicon.com. 2

[4] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 696–712. Springer, 2020. 1

[5] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1175–1186, 2019. 2

[6] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2, 3

[7] K Somani Arun, Thomas S Huang, and Steven D Blostein. Least-squares fitting of two 3-d point sets. *IEEE Transactions on pattern analysis and machine intelligence*, (5):698–700, 1987. 5

[8] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 2

[9] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research*, 36(10):1053–1072, 2017. 3, 7

[10] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE international conference on computer vision*, pages 2300–2308, 2015. 2

[11] Federica Bogo, Javier Romero, Gerard Pons-Moll, and Michael J Black. Dynamic faust: Registering human bodies in motion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6233–6242, 2017. 2

[12] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The euroc micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016. 3

[13] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021. 2, 3, 6, 7

[14] Nicholas Carlevaris-Bianco, Arash K Ushani, and Ryan M Eustice. University of michigan north campus long-term vision and lidar dataset. *The International Journal of Robotics Research*, 35(9):1023–1035, 2016. 3

[15] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM transactions on graphics (TOG)*, 22(3):569–577, 2003. 3

[16] Simone Ceriani, Giulio Fontana, Alessandro Giusti, Daniele Marzorati, Matteo Matteucci, Davide Migliore, Davide Rizzi, Domenico G Sorrenti, and Pierluigi Taddei. Rawseeds ground truth collection systems for indoor self-localization and mapping. *Autonomous Robots*, 27(4):353–371, 2009. 3

[17] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):1–13, 2015. 2

[18] Santiago Cortes, Arno Solin, Esa Rahtu, and Juho Kannala. Advio: An authentic dataset for visual-inertial odometry. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3, 4

[19] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the uzh-fpv drone racing dataset. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6713–6719, 2019. 3

[20] Jeffrey Delmerico and Davide Scaramuzza. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2502–2509. IEEE, 2018. 7

[21] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 2

[22] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *CoRR*, abs/1607.02565, 2016. 3

[23] Marsel Faizullin, Anastasiia Kornilova, Azat Akhmetyanov, and Gonzalo Ferrer. Twist-n-sync: Software clock synchronization with microseconds accuracy using mems-gyroscopes. *Sensors*, 21(1):68, 2021. 4, 6

[24] Marsel Faizullin, Anastasiia Kornilova, Azat Akhmetyanov, Konstantin Pakulev, Andrey Sadkov, and Gonzalo Ferrer. Synchronized smartphone video recording system of depth and rgb image frames with sub-millisecond precision, 2021. 4

[25] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014. 3

[26] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2017. 7

[27] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 1, 3

[28] Chen Gao, Yichang Shih, Wei-Sheng Lai, Chia-Kai Liang, and Jia-Bin Huang. Portrait neural radiance fields from a single image. *arXiv preprint arXiv:2012.05903*, 2020. 2, 3

[29] Xiang Gao, Rui Wang, Nikolaus Demmel, and Daniel Cremers. Ldso: Direct sparse odometry with loop closure. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2198–2204. IEEE, 2018. 3, 7

[30] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[31] Patrick Geneva, Kevin Eckenhoff, Woosik Lee, Yulin Yang, and Guoquan Huang. OpenVINS: A research platform for visual-inertial estimation. In *Proc. of the IEEE International Conference on Robotics and Automation*, Paris, France, 2020. 3, 7

[32] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (TOG)*, 38(6):1–19, 2019. 2

[33] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 2

[34] Sungjin Hong and Yejin Kim. Dynamic pose estimation using multiple rgb-d cameras. *Sensors*, 18(11):3865, 2018. 2

[35] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[36] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12753–12762, 2021. 3

[37] Jinwoo Jeon, Sungwook Jung, Eungchang Lee, Duckyu Choi, and Hyun Myung. Run your visual-inertial odometry on nvidia jetson: Benchmark tests on a micro aerial vehicle. *IEEE Robotics and Automation Letters*, 6(3):5332–5339, 2021. 3, 7

[38] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[39] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[40] Anastasiia Kornilova and Gonzalo Ferrer. Be your own benchmark: No-reference trajectory metric on registered point clouds. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–8, 2021. 2, 5

[41] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual-inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34, 02 2014. 2, 3, 7

[42] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T Barron, and Gleb Gusev. 3d self-portraits. *ACM Transactions on Graphics (TOG)*, 32(6):1–9, 2013. 2

[43] Jinyu Li, Bangbang Yang, Kai Huang, Guofeng Zhang, and Hujun Bao. Robust and efficient visual-inertial odometry with multi-plane priors. In *Pattern Recognition and Computer Vision - Second Chinese Conference, PRCV 2019, Xi'an, China, November 8-11, 2019, Proceedings, Part III*, volume 11859 of *Lecture Notes in Computer Science*, pages 283–295. Springer, 2019. 3, 7

[44] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T Freeman. Learning the depths of moving people by watching frozen people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4521–4530, 2019. 2

[45] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405*, 2021. 3

[46] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), July 2019. 1, 3

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 3

[48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 7

[50] Anastasios I. Mourikis and Stergios I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, 2007. 7

[51] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015. 3

[52] Raul Mur-Artal and Juan D Tardós. ORB-SLAM2: An open-source slam system for monocular, stereo, and rgb-d

cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 3, 6

[53] Janosch Nikolic, Michael Burri, Igor Gilitschenski, Juan Nieto, and Roland Siegwart. Non-parametric extrinsic and intrinsic calibration of visual-inertial sensor systems. *IEEE Sensors Journal*, 16(13):5433–5443, 2016. 4

[54] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407. IEEE, 2011. 4

[55] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 international conference on 3D vision (3DV)*, pages 484–494. IEEE, 2018. 3

[56] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[57] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 3

[58] Bernd Pfrommer, Nitin Sanket, Kostas Daniilidis, and Jonas Cleveland. Penncosyvio: A challenging visual inertial odometry benchmark. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3847–3854, 2017. 3, 4

[59] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2242–2251, 2019. 2

[60] Tong Qin, Shaozu Cao, Jie Pan, and Shaojie Shen. A general optimization-based framework for global pose estimation with multiple sensors, 2019. 7

[61] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 3, 7

[62] Tong Qin, Jie Pan, Shaozu Cao, and Shaojie Shen. A general optimization-based framework for local odometry estimation with multiple sensors, 2019. 3, 7

[63] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020. 5

[64] Jan Razlaw, David Droeschel, Dirk Holz, and Sven Behnke. Evaluation of registration methods for sparse 3d laser scans. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–7. IEEE, 2015. 5

[65] Joern Rehder, Janosch Nikolic, Thomas Schneider, Timo Hinzmann, and Roland Siegwart. Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4304–4311. IEEE, 2016. 4

[66] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, 2020. 2, 7

[67] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021. 2, 7

[68] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020. 2, 3, 7

[69] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 3

[70] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 3

[71] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 6, 7

[72] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 7

[73] David Schubert, Thore Goll, Nikolaus Demmel, Vladyslav Usenko, Jörg Stückler, and Daniel Cremers. The tum vi benchmark for evaluating visual-inertial odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1680–1687, 2018. 3, 4

[74] Artem Sevastopolsky, Savva Ignatiev, Gonzalo Ferrer, Evgeny Burnaev, and Victor Lempitsky. Relightable 3d head portraits from a smartphone video. *arXiv preprint arXiv:2012.09963*, 2020. 2, 3

[75] Ari Shapiro, Andrew Feng, Ruizhe Wang, Hao Li, Mark Bolas, Gerard Medioni, and Evan Suma. Rapid avatar capture and simulation using commodity depth sensors. *Computer Animation and Virtual Worlds*, 25(3-4):201–211, 2014. 2

[76] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2387–2397, 2019. 1

[77] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 2

[78] STMicroelectronics. *iNEMO inertial module: always-on 3D accelerometer and 3D gyroscope*, 2019. Rev. 2. 4

[79] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012. 3

[80] Shinya Sumikura, Mikiya Shibuya, and Ken Sakurada. OpenVSLAM: A Versatile Visual SLAM Framework. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 2292–2295, New York, NY, USA, 2019. ACM. 2, 3, 7

[81] Garvita Tiwari, Bharat Lal Bhatnagar, Tony Tung, and Gerard Pons-Moll. Sizer: A dataset and model for parsing 3d clothing and learning size sensitive 3d clothing. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 2

[82] George Toderici, Georgios Evangelopoulos, Tianhong Fang, Theoharis Theoharis, and Ioannis A Kakadiaris. Uhdb11 database for 3d-2d face recognition. In *Pacific-Rim Symposium on Image and Video Technology*, pages 73–86. Springer, 2013. 2

[83] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, pages 501–510, 2019. 2

[84] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 1991. 5

[85] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–11. 2009. 2

[86] Cheng Wang, Yu Zhao, Jiabin Guo, Ling Pei, Yue Wang, and Haiwei Liu. NEAR: The NetEase AR oriented visual inertial dataset. In *International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 366–371. IEEE, 2019. 3

[87] Ziyan Wang, Timur Bagautdinov, Stephen Lombardi, Tomas Simon, Jason Saragih, Jessica Hodgins, and Michael Zollhofer. Learning compositional radiance fields of dynamic human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5704–5713, 2021. 1, 2, 3

[88] Qingshan Xu and Wenbing Tao. Planar prior assisted patchmatch multi-view stereo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12516–12523, 2020. 1, 7

[89] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018. 2

[90] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of human body shape in motion with wide clothing. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 439–454, Cham, 2016. Springer International Publishing. 2

[91] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 2

[92] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020. 2

[93] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 1

[94] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1

[95] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017. 2

[96] Licong Zhang, Jürgen Sturm, Daniel Cremers, and Dongheui Lee. Real-time human motion tracking using multiple depth cameras. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2389–2395. IEEE, 2012. 2

[97] Shishun Zhang, Longyu Zheng, and Wenbing Tao. Survey and evaluation of rgb-d slam. *IEEE Access*, 9:21367–21387, 2021. 6

[98] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5560–5569, 2021. 2

[99] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. *arXiv preprint arXiv:2105.00261*, 2021. 2

[100] David Zuñiga-Noël, Alberto Jaenal, Ruben Gomez-Ojeda, and Javier Gonzalez-Jimenez. The uma-vi dataset: Visual–inertial odometry in low-textured and dynamic illumination environments. *The International Journal of Robotics Research*, 39(9):1052–1060, 2020. 3