A survey on RGB-D datasets

Alexandre Lopes\textsuperscript{a,**}, Roberto Souza\textsuperscript{b,c}, Helio Pedrini\textsuperscript{a}

\textsuperscript{a}Institute of Computing, University of Campinas, Brazil  
\textsuperscript{b}Department of Electrical and Computer Engineering, University of Calgary, Canada  
\textsuperscript{c}Hotchkiss Brain Institute, University of Calgary, Canada

ABSTRACT

RGB-D data is essential for solving many problems in computer vision. Hundreds of public RGB-D datasets containing various scenes, such as indoor, outdoor, aerial, driving, and medical, have been proposed. These datasets are useful for different applications and are fundamental for addressing classic computer vision tasks, such as monocular depth estimation. This paper reviewed and categorized image datasets that include depth information. We gathered 231 datasets that contain accessible data and grouped them into three categories: scene/objects, body, and medical. We also provided an overview of the different types of sensors, depth applications, and we examined trends and future directions of the usage and creation of datasets containing depth data, and how they can be applied to investigate the development of generalizable machine learning models in the monocular depth estimation field.

© 2022 Elsevier Ltd. All rights reserved.

1. Introduction

Depth is a critical information for many computer vision and image analysis applications. For example, it has been applied for tasks such as synthetic object insertion in computer graphics (Luo et al., 2020), robotic grasping (Lenz et al., 2015) automatic 2D to 3D conversion in film (Xie et al., 2016), robot-assisted surgery (Stoyanov et al., 2010), and autonomous driving (Levinson et al., 2011).

Despite using depth sensors that capture the distance information, researchers also use stereo vision matching to infer it, especially for its condensed size and cost. Lately, deep learning methods are being used to produce more precise and dense depth maps. For example, they can improve finer-grained details (Miangoleh et al., 2021), produce dense maps from sparse inputs (Uhlig et al., 2017), and refine depth for mirror surfaces (Tan et al., 2021a).

An important field of study for depth is monocular depth estimation, especially because it does not require using depth sensors, reducing the size and cost of computer vision systems’ setups. Also, it can be applied to existing monocular systems, that comprise the majority of image capturing systems available. For instance, Light Detection And Ranging (LiDAR) scanners usually cost thousands of dollars, and their cost and weight can be impractical for many small drone applications.

As a result of the extensive range of applications of depth, a considerable number of datasets include distance measurements of points of the scene they acquire. These datasets are collected using different sensors in distinct scenes for applications such as Simultaneous Localization and Mapping (SLAM) (Sturm et al., 2012), Reconstruction (Dai et al., 2017), Object Segmentation (McCormac et al., 2017), and Human Activity Recognition (Zhang et al., 2016a). With the increasing number and diversity of datasets, researchers were able to explore more generalistic forms of depth estimation, leading to techniques focused on zero-shot cross-dataset depth estimation (Li and Snavely, 2018a; Xian et al., 2020a; Ranftl et al., 2021a, 2020). The idea is to produce powerful methods able to estimate depth for in-the-wild scenes, increasing the range of applications for depth estimation.

The main contribution of this paper is to categorize and summarize the existing datasets with depth data. We propose a survey that can be used by researchers of both individual applications and general systems. While there are good reviews of RGB-D datasets (Firman, 2016; Cai et al., 2017), the most
recent one was published in 2017, and datasets have evolved both in complexity and size since then. Our survey presents a comprehensive literature review on more than 200 publicly available datasets included from an initial list of more than 300 datasets. Nearly half of the public datasets were published in 2017 or after, therefore, not included in any other review. We also made this work available on a website to facilitate the filtering by application, scene type, sensor, and year.

The remainder of this paper is structured as follows. In the next section, we discuss and categorize depth sensors, explaining the main differences and applications for each category. In Section 2.1, we present the methodology used to perform the literature review. In Section 2.3, we present the datasets divided into categories, describing the most influential datasets for each category and presenting the rest in tables. In Section 2.5, we present tendencies and discuss future directions for RGB-D data usage. Finally, we provide a summary of the field and discuss how the area is evolving in Section 6.

2. Sensors

Range (or depth) data is crucial for understanding the 3D scene projected onto a 2D plane forming an image. There are multiple ways to obtain such information, either using a depth sensor or estimating depth. A depth sensor is a device that provides the distance from the sensor to an element in the scene, although it is possible to collect distance information using two or more RGB cameras from a scene. We define as Stereo Camera Sensing, all systems formed by two or more cameras. Therefore, light field cameras are also included here.

Previously, authors proposed distinct divisions for the types of sensors (Fisher and Konolige, 2008; Choi, 2019). In this survey, we use a categorization of depth sensors inspired by Choi (2019)’s work. We divide the sensors into the following categories: Structured Light, Time-of-Flight (TOF), Light Detection and Ranging (LiDAR), and Stereo Camera Sensing. We display examples of each category in Figure 1.

Ultrasonic and Radar sensors also produce distance information, but they are out of the scope of this work because they are rarely used to produce depth information associated with RGB data. We detail each one of the sensors categories in the following sub-sections and show the differences of these types and possible application scenes in Table 1.

2.1. Structured Light

Structured Light sensors (also called Active Stereo sensors) rely on a projector of light captured by a camera. The simplest way to achieve such a goal is to project a point with a device and capture this point in the scene with the camera. The depth of this point can be measured by a technique called Triangulation. For estimating depth, it is necessary to find the position of the projected point in the image plane, have the distance between the camera and the light projector, the camera’s internal parameters, and the position in space of the projector. With this information, it is possible to create a triangle and calculate the height of the triangle formed by the camera, projector, and illuminated scene point to determine the distance. The strategy of projecting points would be slow in practice since it is necessary to project a point for every position that is represented as a pixel in the image.

A more efficient strategy is to project the light as a stripe that associated with different coding strategies, such as the Binary Coded Structured Light strategy, can reduce the number of frames necessary to produce a full depth map. It can also be coded with RGB lights. Details about different codification strategies are discussed by Salvi et al. (2004).

Most Structured Light sensors do not work under direct sunlight since they rely on light projection in a scene. Therefore, they are usually suitable for indoor scene applications. Researchers have proposed strategies to overcome challenging light conditions (O’Toole et al., 2015), and new sensors appear in smartphones for face identification systems for both indoor and outdoor scenes. They typically have a low range limit, not going further than 10 meters. Examples of this type of sensor include Matterport, Kinect v1, and RealSense SR300 cameras.

2.2. Time-of-Flight

TOF sensors estimate the distance of an object in the scene to a sensor by measuring the time it takes for an emitted light to be received by the sensor. Therefore, TOF sensors rely on the time that a light wave takes to go to a point in a scene and to be reflected to a sensor. The concept is barely the same as the Ultrasonic and Radar Sensors, but here light is used as the emitted signal.

There are multiple strategies for capturing the time-of-flight of light. The most straightforward strategy is using a technique called Pulse Modulation, where a very fast pulse of light is emitted and then received by the sensor. The time delay between the emitted light pulse and the received light pulse is used to compute the distance of the object in the scene. Continuous-Wave Modulation is another strategy, where the light is modulated by its intensity, and the distance is measured by calculating the shift in phase of the original emitted light and the received light.

TOF sensors generally are compromised under strong sunlight conditions (Kazmi et al., 2012), making this sensor more commonly applied to indoor scenes. Existing studies try to overcome the effect under intense background light (Buttgen and Seitz, 2008) and to reduce the measurement uncertainty under such conditions. Examples of this type of sensor include Kinect v2 (Xbox One sensor), SoftKinetic DS 325, and RIEGL VZ-400.

2.3. LiDAR

LiDAR sensors use the same idea of measuring the time that an emitted light is received by a sensor, but they rely on one or multiple laser beams (concentrated light) to produce depth measurements of points in the scene, and the device usually has a rotating mirror to generate 360° scans of a scene. Hence, LiDAR sensors produce point clouds of a scene, not a dense
depth map of it. They rely on focused laser beams, which allow them to collect distance measurements as far as a few kilometers. LiDAR sensor models have different specifications (e.g., resolution, scans per second, and distance accuracy), and some scans are built in a multilayer (multiple laser beams) configuration, allowing them to measure not only in a 360° plane of the sensor but in 3D.

LiDAR measurement accuracy is usually independent of distance, although some models can fail in adverse weather conditions, such as dense fogs and turbulent snow (Jokela et al., 2019). Each LiDAR point also includes the intensity measurements, which can be interpreted as a measurement of reflectivity of the point that the light hit. This value is suitable for many applications, such as vegetation cover understanding and tunnel damage detection (Kashani et al., 2015), giving LiDAR additional information that other types of sensors do not produce.

LiDAR sensors emit light; therefore, they work in difficult lighting conditions, such as dark environments. They are suitable for indoor and outdoor application scenes, but the available models are usually limited to specific applications, such as aerial measurements, outdoor/driving applications, and small indoor spaces depth estimation. Examples of such types of sensors include Velodyne Sensors, Faro Focus 3D Laser, and SICK LMS-511.

2.4. Stereo Camera Sensing

We define here Stereo Camera Sensing (SCS) as any system formed by two or more image sensors or lenses used to produce a Depth Map of a scene. Hence, simplistic pairs of cameras and complex light field systems composed by multiple microlenses are both identified in the same category. A straightforward strategy to measure depth from two or more cameras is Triangulation. The Triangulation idea is the same as applied in Structured Light sensors, but using a camera instead of a projector. The idea is that finding the position of a pixel in the image plane of camera A projected from a point P in the space, and the position of a pixel projected by the same point P in camera B, it is possible to find the depth of that point in a scene with the intrinsic parameters of the camera. After finding both lines in camera B, it is only necessary to know the distance between the two cameras (baseline distance) and internal parameters of the cameras to know the depth of the point P.

A limitation of this strategy occurs when the point of interest has no texture. For instance, it is practically impossible to determine which point of a smooth painted wall observed in the image projected by camera A is equivalent to the image projected by camera B. Therefore, it is difficult to determine a point’s depth with acceptable accuracy without the correspondence of the pixels in both image planes. Recently, Deep Learning based methods have tried to address this limitation, increasing the accuracy of the estimation (Zbontar and LeCun, 2015). Examples of such types of sensors include light field cameras and ZED cameras.

3. Methodology

A literature review should synthesize previous knowledge, identify biases and gaps in the literature (Rowe, 2014). Since our study aims to describe, categorize, and identify future trends for RGB-D datasets, we defined a non-conventional methodology to find the related papers. Instead of defining search terms to find the papers directly, we collected datasets using backward snowballing. The premise is that many datasets containing depth data do not have depth estimation as their primary goal, as in KITTI Dataset (Geiger et al., 2013). Therefore, defining search strings that could find depth datasets using generalist terms would result in numerous false-positive results. For instance, the search string RGB-D OR Depth AND Dataset searching in abstract, keywords, or title brings more than 23 thousand results in Scopus. Moreover, if we define a complex composed search string to filter the results, we would miss many datasets in the search.

As monocular depth estimation, salient object detection, and action recognition are prominent fields in the area, we defined the following search string to perform backward snowballing: ("single image" OR monocular) AND depth AND estimation) OR ("Salient Object Detection" OR "Action Recognition") AND RGB-D). The terms “monocular” and “single image” are applied mainly for monocular depth estimation but are also used for stereo trained systems, depth completion, and other applications. We conducted the review in Scopus and Google Scholar search engines. In Scopus, we revised all papers from January 1st, 2016, through August 31st, 2021. From Google Scholar, we followed the same dates, but we also included a stop criterion. If we found one search page without relevant items, we would end the year’s search. The inclusion of Google Scholar is justified because many relevant papers are published in arXiv. Consequently, those could also be included in this work.

The exclusion and inclusion criteria for papers are defined in Table 2. These criteria are applied to the papers found using the previous search term. After excluding papers, backward snowballing was applied to find the datasets used/described by the remaining works. Initially, we reviewed 2,119 papers, which led to 374 dataset candidates. We also applied an exclusion criterion to these candidates, and only papers with active project
Fig. 1: Examples of depth data with image (first row) and depth (second row) of the following sensors: (a) Structured Light from NYUv2 [Silberman et al., 2012], (b) TOF from AVD [Ammirato et al., 2017], (c) LiDAR from KITTI [Geiger et al., 2013], and (d) Stereo Camera Sensing from ReDWeb [Xian et al., 2018], where the authors compute correspondence maps by using optical flow.

Table 2: Inclusion and Exclusion Criteria.

| Criterion                                      | Category   |
|------------------------------------------------|------------|
| Papers that discuss depth estimation           | Inclusion  |
| Papers using depth sensors, stereo image      | Inclusion  |
| sensing, or synthetic data                    |            |
| Papers not written in English                 | Exclusion  |
| Papers exclusively using private datasets     | Exclusion  |
| Papers not presenting minimal evidence of valid results | Exclusion   |
| Duplicated paper/report. We kept the most complete one | Exclusion   |

websites, contact information to download the dataset, or direct download link were included. Hence, the final list of datasets to be included was reduced to 231 datasets.

4. Datasets

In recent years, many datasets have been created using the sensors or stereo vision sensing presented in the previous section. In addition to datasets using real data, this paper also includes datasets containing synthetic data. These were created mainly by simulation systems and often presented extra data such as semantic segmentation and 3D object detection bounding boxes. We divided the selected datasets into three different categories and six different sub-categories representing different application areas. The taxonomy tree is available in Figure 2

The categories represent the intended application of the dataset. In the first level, we identify datasets that are mainly interested in Scenes/Objects, Human Body, or Medical Applications. The following sub-sections explore each application area, and list all of them in each sub-category’s table. We also detail three, two or one datasets for each sub-category, based on the total number of datasets of each sub-category. If we detail three papers, the two first ones are the most cited papers that contain complementary scenarios. For example, KITTI Dataset and ScanNet Dataset contain street and indoor scenes, respectively. The third paper is the most cited paper published in 2017 or later. If we detail two papers, these are the most cited ones that contain complementary scenarios, and if we detailed one paper, it is the most cited in the sub-category.

4.1. Scene/Objects

In this category, we grouped all datasets generally intended to expose scenes, individual objects, or groups of objects containing or not humans. Therefore, datasets that reconstruct scenes/objects, segment elements of a scene, salient objects using depth, and contain exclusively depth maps are sub-categorized here. We created an “Other” sub-category to accommodate datasets that did not fit into these previous sub-categories.

Some papers explore multiple applications, primarily synthetic datasets, since they can create reconstruction and segmentation data directly using simulation environments. These papers are presented in one of their application areas to reduce redundancy. The only exception is for datasets of “SLAM, Odometry, or Reconstruction” and “Segmentation or Other Extra Information” sub-categories that are presented together in Table 5 since this combination is very frequent for datasets.

4.1.1. SLAM, Odometry, or Reconstruction

This sub-category contains multiple types of applications, however, all of them have a common characteristic: they present extra information that makes possible to recreate in any detail level, a 3D scene. For SLAM and odometry related papers, they typically present camera pose information, giving position and orientation of the capturing apparatus of each frame/image. We treated odometry differently from SLAM since odometry essentially aims to estimate the path of the camera, and SLAM tries to obtain a consistent trajectory and scene map of the camera [Yousif et al., 2015].

All collected datasets that contain data exclusively for SLAM, Odometry, or Reconstruction are shown in Table 8. In general, applications of indoor scenes focus on reconstruction, and external scenes (such as driving scenes) focus on SLAM/odometry. Table 8 also contains datasets of this sub-category, however, with extra annotated information such as semantic segmentation data. Some of the most cited datasets in the field include:

KITTI Dataset. Analyzing the datasets presented in this paper, this is the most cited one. The KITTI Dataset consists of a complex system of IMU/GPS, LiDAR scanner, and multiple
cameras (Geiger et al., 2013). They recorded 6 hours of traffic scenes and, in addition to collecting the information from the sensors, provided data from 3D object detection bounding boxes, optical flow, and visual odometry/SLAM (Geiger et al., 2012). The project was expanded over the years, and the authors included data for tracking, road/lane detection, semantic/instance segmentation, and depth completion. Its depth completion data is composed of 94 thousand depth annotated RGB images (Uhrig et al., 2017) to produce dense depth maps from LiDAR points.

This dataset influenced the creation of the synthetic datasets Virtual KITTI (Gaidon et al., 2016) and Virtual KITTI 2 (Cabon et al., 2020). Recently, the KITTI authors released the KITTI-360 Dataset (Liao et al., 2021), which has more cameras, sensors, and more annotated data than the original KITTI Dataset.

**ScanNet Dataset.** ScanNet is an indoor dataset collected using an occipital structure sensor - a structured light sensor similar to Microsoft Kinect v1 (Dai et al., 2017). The authors performed a dense reconstruction and conducted object instance-level annotation of all surfaces in the reconstruction. They also conducted a CAD Model Retrieval and Alignment for the objects in the scenes, which means that a 3D CAD model represented each instance of the annotated object in a scene. This dataset contains 2.5M views in 2,119 different scenes.

**SunCG Dataset.** The project associated with this dataset is focused on semantic scene completion, where from a single point of view, it estimates a complete 3D representation with the semantic label associated with the scene (Song et al., 2017). Instead of estimating the semantic segmentation of visible surfaces, this project aims to predict the occluded space (3D scene representation) and a label for each voxel in the scene. Therefore, it deals with Reconstruction and Segmentation as a unified task. This dataset comprises synthetic data containing an entire 3D model scene (which can be related to reconstruction), with semantic labels associated with it.

### 4.1.2. Segmentation or Other Extra Information

In this sub-category, all datasets have extra information that leads to a better scene understanding. Extra information can be seen as semantic or instance segmentation, 2D or 3D object detection, optical flow, salient object detection, etc. For instance, datasets that explore potential applications for depth estimation algorithms and semantic segmentation, and datasets dedicated to salient object detection were categorized here.

The complete list of datasets containing extra information is available in Table 4. We provide the type of extra information for each dataset in the “Extra Data” column. Researchers interested in a specific application, for instance, salient object detection, should use it to filter datasets related to their field of interest. Table 5 also reports datasets for this sub-category, as well as information of “SLAM, Odometry, or Reconstruction” sub-category. Therefore, researchers interested in semantic segmentation datasets may check both tables and refer to the “Extra Data” column to find the datasets that match their interest. Next, three of the most influencing and promising papers for this sub-category are presented.

**NYUv2.** This dataset contains indoor images and is the most cited dataset for this type of scene in the “Segmentation or Other Extra Information” sub-category. It was collected using Microsoft Kinect v1 sensor and is composed of aligned RGB and depth images, labeled data containing semantic segmentation, and raw data (Silberman et al., 2012). This project is a continuation of NYUv1 (Silberman and Fergus, 2011), which uses the same sensor and type of data, but has fewer scenes and total frames.

**Scene Flow Datasets.** This dataset is a collection of three datasets: FlyingThing3D, Monkaa, and Driving. The first is composed of everyday objects flying along random
| Dataset Name | Ref. | Year | Scene Type | Sensor Type | Sensor Name | Data Modalities | Extra Data | Images/Scenes |
|--------------|------|------|------------|-------------|-------------|----------------|------------|---------------|
| GL3D         | Shen et al. 2018 | 2018 | Aerial     | SCS         | Stereo Camera | Color, Depth | -           | 543 Scenes (125623 Images) |
| ApolloScape  | Wang et al. 2019 | 2020 | Driving    | LiDAR       | Velodyne HDL-64E | Color, Depth, GPS, Radar | -           | 155 Min With 93k Frames |
| KAIST        | Jeong et al. 2019 | 2019 | Driving    | SCS, LiDAR  | Velodyne VLP-16, SICK LMS-511, Stereo Camera | Color, Depth, GPS, IMU, Altimeter | -           | 19 Sequences (191 Km) |
| RobotCar     | Maddern et al. 2017 | 2016 | Driving    | LiDAR       | 2 X SICK LMS-151 2D LiDAR, 1 X SICK LD-MRS 3D LiDAR | Color, Depth, GPS, INS (Inertial Navigation System) | -           | 133 Scenes (almost 20M Images (from Multiple Sensors)) |
| Malaga Urban | Blance-Claraco et al. 2014 | 2014 | Driving    | SCS, LiDAR  | Velodyne HDL-64E, Stereo Camera | Color, Depth, IMU, GPS | -           | 15 Sequences |
| Omnidirectional | Schönberg et al. 2014 | 2014 | Driving    | SCS, LiDAR  | Velodyne HDL-64E, Stereo Camera | Color, Depth | -           | 152 Scenes (12607 Frames) |
| Ford Campus Vision And LiDAR | Pandej et al. 2011 | 2011 | Driving    | SCS, LiDAR  | Velodyne HDL-64E, Stereo Camera | Color, Depth, IMU, GPS | -           | 2 Sequences |
| Karlshue     | Geiger et al. 2011 | 2011 | Driving    | SCS         | Stereo Camera | Color, GPS/IMU | -           | 20 Sequences (16657 Frames) |
| Multi-FoV (Urban Canyon) | Zhang et al. 2018 | 2016 | Driving, Indoor | - | Synthetic | Color, Depth | -           | 2 Sequences |
| – Zessel et al. 2013 | 2013 | Driving, Outdoor | N/A | RGB-D Scans (N/A) | Color, Depth | -           | 13 Scenes (5 Castle, 5 Church, 3 Street Scenes) |
| BlendedStereo Camera | Yao et al. 2020 | 2020 | In-the-wild | - | Synthetic | Color, Depth | -           | - |
| Youtube3D    | Chen et al. 2019 | 2019 | In-the-wild | - | Two Pouns Automatically Annotated | Color, Relative Depth | -           | 795066 Images |
| – Malteson et al. 2019 | 2019 | In-the-wild | Structured Light, TOF | Kinect V1, V2 And Synthetic | Color, Depth | -           | 10 Sequences (2703 Frames) |
| 4D Light Field Benchmark Habitat | Honauer et al. 2016 | 2016 | In-the-wild | - | Light-field (Synthetic) | Color, Depth | -           | 24 Scenes |
| Habitat Matterport (HMDJ) | Ramakrishnan et al. 2021 | 2021 | Indoor | Structured Light | Matterport Pno2 | Color, Depth | -           | 1000 Scenes |
| MilliEgo      | Lai et al. 2020 | 2020 | Indoor     | Structured Light, LiDAR | Intel D435I Depth, Velodyne HDL-32E, Velodyne Ultra Puck | Color, Depth, IMU, LiDAR, Depth | -           | 17 Distinct Floors From 6 Different Multistory |
| ODS          | Lai et al. 2019 | 2019 | Indoor     | SCS         | MiniPolar 360 Camera (Stereo Camera) | Color, Depth | Normal Maps | 6 Indoor Areas (50k Images) |
| 360D         | Goullis et al. 2018 | 2018 | Indoor     | Structured Light | Synthetic And Matterport Camera | Color, Depth | -           | 12072 Scanned Scenes And 10024 CG Scenes |
| PanoSUNCG    | Wang et al. 2018 | 2018 | Indoor     | - | Synthetic | Color, Depth | -           | 103 Scenes (25k Images) |
| CoRBS        | Wassenmüller et al. 2016 | 2016 | Indoor     | TOF         | Kinect V2 | Color, Depth | -           | 4 Scenes (9 Hours Of Recording) |
| EuRoC MAV    | Baur et al. 2016 | 2016 | Indoor     | TOF, Stereo Camera | Vision Motion Capture, Leica MS50 | Color, Depth, IMU | -           | 11 Scenes |
| Augmented ICL-NUM | Choi et al. 2015 | 2015 | Indoor     | - | Synthetic | Color, Depth | -           | 4 Scenes (2 Living Room, 2 Offices) |
| Ikea         | Li et al. 2015 | 2015 | Indoor     | Structural Light | Kinect V1 And PrimeSense | Color, Depth | -           | 7 Scenes |
| ViBRIO       | Martínez-Gómez et al. 2013 | 2013 | Indoor     | Structural Light | Kinect V1 | Color, Depth | Semantic Category of the Scene | 5 Sequences (22454 Images) |
| ICL-NUM      | Hamba et al. 2014 | 2014 | Indoor     | - | Synthetic | Color, Depth | -           | 8 Scenes (4 Living Room, 4 Office) |
| MobileRGBD   | Vanfleteren and Negre 2014 | 2014 | Indoor | TOF | Kinect V2 | Color, Depth | -           | 3 Scenes (9.5 Hours Of Recording) |
| RGBD Object V2 | Lai et al. 2014 | 2014 | Indoor     | Structured Light | Kinect V1 | Color, Depth | -           | 14 Sequences |
| – Mattausch et al. 2014 | 2014 | Indoor | LiDAR | Faro Focus 3D Laser | Depth | -           | 40 Scenes (rooms from Three Offices) |
| RGB-D 7 Scenes | Blocher et al. 2013 | 2013 | Indoor     | Structural Light | Kinect V1 | Color, Depth | -           | 7 Scenes (500-1000 Frames/scene) |
| Reading Room | Zhou et al. 2013 | 2013 | Indoor     | Structural Light | Asus Xion Pro Live | Color, Depth | -           | 1 Scene |
| TUM-RGBD     | Srinath et al. 2012 | 2012 | Indoor     | Structural Light | Kinect V1 | Color, Depth, Accelerometer | -           | 39 Sequences |
| IROS 2011 Paper Kinect | Ponteufeas et al. 2011 | 2011 | Indoor     | Structural Light | Kinect V1 | Depth | -           | 27 Sequences |
| – Zhou and Krumm 2013 | 2013 | Indoor, Isolated Objects / Focussed On Objects | Structured Light | Asus Xion Pro Live | Color, Depth | -           | 6 Scenes |
| – Meister et al. 2012 | 2012 | Indoor, Isolated Objects / Focussed On Objects | Structured Light, TOF | Kinect Fusion (Kinect V1) For Two Scenes. Rigl V2-400 For Office | Color, Depth | -           | 2 Scenes: Statue And Targetbox |
| M&M          | Ken et al. 2020 | 2020 | Indoor, Outdoor | SCS | Stereo Camera | Color, Depth | -           | 46900 Scenes (170k Frames) And 130k Images |
| Manequin Challenge | Li et al. 2019 | 2019 | Indoor, Outdoor | SCS | Stereo Camera | Color | -           | 46900 Scenes (170k Frames) |
| Stereo CameraIC | Zhu et al. 2018 | 2018 | Indoor, Outdoor | SCS, LiDAR | Velodyne (LiDAR), Stereo Camera | Color, Depth, IMU | -           | 5 Sequences |
| ETH3D        | Schöps et al. 2017 | 2017 | Indoor, Outdoor | SCS, LiDAR | Stereo Camera | Color, Depth | -           | 25 High-res, 10 Low-res |

Continue on Next Page
Trajectories [Mayer et al., 2016]. The second was created using Blender computer graphics software, based on the information from an animated short film called Monkaa. The third is composed of a street scene. Scene Flow contains only synthetic data for all three datasets and, in addition to depth and RGB frames, the authors also include optical flow, segmentation, and stereo disparity change data.

Waymo Perception. This dataset is a street scene dataset composed of RGB and LiDAR labels. It consists of street scenes, and the authors labeled LiDAR using 3D bounding boxes for vehicles, pedestrians, cyclists, and signs [Sun et al., 2020]. They also provide RGB images annotations with 2D bounding boxes of vehicles, pedestrians, and cyclists. The 3D bounding boxes also have unique tracking IDs for tracking applications. The Waymo Perception Dataset is composed of 1,150 scenes with 20 seconds of recording each.

4.1.3. Depth Data Only

The datasets presented here are for the specific purpose of training depth estimation algorithms. They do not directly provide reconstruction, SLAM, or other information, although some of these applications are direct results of depth estimation. For example, these works explore monocular depth estimation [Cho et al., 2021b], zero-shot depth estimation [Yin et al., 2020], and multi-camera depth estimation [Antequera et al., 2020].

We present all papers found specifically for depth estimation in Table 6. All datasets for all categories and sub-categories in this paper also contain depth information as it is an inclusion criterion for papers to be incorporated to this work. Some relevant papers in this sub-category are:

**RedWeb Dataset.** This dataset deals with the in-the-wild scenario, covering scenes such as street, office, park, farm, etc. As formed in the acronym of this dataset’s name “Relative Depth from Web” (RedWeb), this dataset is formed by stereo images collected from the Internet [Xian et al., 2018]. The authors use optical flow to generate correspondence maps and create a relative depth map of the image. They post-process the data by segmenting the sky to increase the quality of the depth maps.

**SQUID Dataset.** This dataset is composed of underwater images collected from four different sites: two in the Red Sea and two in the Mediterranean Sea [Berman et al., 2021]. In addition to collecting stereo pair images, the authors included a ColorChecker to propose color restoration techniques in underwater images.

**Middlebury Datasets.** These datasets are a composition of data released in different papers over the years of 2001, 2003, 2005, 2006, and 2014. These datasets are acquired using different strategies: custom structured light using a video projector for the Middlebury 2003 (Scharstein and Szeliski, 2003), Middlebury 2005 (Scharstein and Pal, 2007), Hirschmüller and Scharstein (2011b), Middlebury 2006 (Scharstein and Pal, 2007, Hirschmüller and Scharstein, 2007), and Middlebury 2014 (Scharstein et al., 2014), while Middlebury 2001 (Scharstein and Szeliski, 2002) uses stereo image pair dis-
Table 4: Datasets of “Segmentation or Other Extra Information”

| Dataset Name       | Ref.            | Year | Scene Type | Application          | Sensor Type | Sensor Name | Data Modalities                        | Extra Data                                                                 | Images/Scenes                  |
|--------------------|-----------------|------|------------|----------------------|-------------|-------------|----------------------------------------|----------------------------------------------------------------------------|-------------------------------|
| VALID              | Chen et al. (2020) | 2020 | Aerial     | SOE                  | -           | Synthetic   | Color, Depth                          | Object Detection, Panoptic Segmentation, Instance Segmentation, Semantic Segmentation | 6 Scenes (6690 Images)         |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| US3D               | Foster et al. (2020) | 2019 | Aerial     | SOE                  | LiDAR       | Airborne LiDAR | Color, Depth                          | Semantic Segmentation                                                     | 4160 Images From 3 Different Cities (a Fourth Is Not Available)            |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Vaihingen           | Nada et al. (2010) | 2011 | Aerial     | SOE                  | LiDAR       | Leica ALS50 And ALTM-ORION M           | Color, Depth                          | Semantic Segmentation                                                      | 33 Patches                     |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Potdam             | Rolfsensier et al. (2012) | 2011 | Aerial     | SOE                  | N/A         | N/A         | Color, Depth                          | Semantic Segmentation                                                      | 38 Patches                     |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Leddar P Iriset     | D'Aniello et al. (2021) | 2021 | Driving    | SOE and Tracking (Other) | LiDAR       | Leddar Ppixell LiDAR | Color, Depth, IMU, Radar              | 3D Bounding Boxes, 2D Bounding Boxes, Semantic Segmentation              | 97 Sequences (20k Frames)     |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Virtual Kitti 2     | Cabou et al. (2020) | 2020 | Driving    | SOE and Tracking (Other) | -           | Synthetic   | Color, Depth                          | Semantic Segmentation, Instance Segmentation, Optical Flow              | 5 Scenes (multiple Conditions For Each Scene)                            |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Waymo Perception    | Sun et al. (2020) | 2020 | Driving    | SOE                  | LiDAR       | N/A         | Color, Depth                          | 3D Object Detection                                                       | 1150 Scenes (20 Seconds/scene)                                         |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Argoverse           | Chong et al. (2019) | 2019 | Driving    | SOE and Tracking (Other) | SCS, LiDAR  | Argo LiDAR, Stereo Camera            | Color, Depth                          | Semantic Segmentation                                                      | 113 Scenes                     |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| CityScapes          | Cordes et al. (2016) | 2016 | Driving    | SOE                  | SCS         | Stereo Camera | Color, Odometry | Segmentation, 3D-object Detection And Pose | 50 Cities (25k Frames)                                                |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| SYNTHIA             | Ros et al. (2016) | 2016 | Driving    | SOE                  | Virtual 8 Depth Sensors | Synthetic   | Color, Depth                          | Instance Segmentation                                                      | 5 Sequences (with Sub-sequences) At 5 Fps. 20k Images From Videos      |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Daimler Urban Segmentation Ground Truth Stixel | Scharwächter et al. (2014) | 2014 | Driving    | SOE                  | SCS         | Stereo Camera | Color                          | Semantic Labeling                                                        | 5k Images                      |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
|                  | Fleisher et al. (2015) | 2013 | Driving    | SOE                  | SCS         | Stereo Camera | Color                          | Stixels                                                                    | 12 Sequences                  |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Daimler Stereo Pedestrian | Kellar et al. (2011) | 2011 | Driving    | SOE                  | SCS         | Stereo Camera | Color                          | Object Detection                                                         | 28919 Images                  |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Unreal             | Manca et al. (2013) | 2018 | Driving, Outdoor | SOE                  | -           | Synthetic   | Color, Depth                          | Semantic Segmentation                                                      | 21 Sequences (100k Images)                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| OASIS V2           | Chen et al. (2020) | 2021 | In-the-wild | SOE                  | -           | From Human Annotation | Color, Depth | Semantic Segmentation, Normal Maps, Instance Segmentation | 102k Images                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| OASIS              | Chen et al. (2020) | 2020 | In-the-wild | SOE                  | -           | From Human Annotation | Color, Depth | Semantic Segmentation, Normal Maps, Instance Segmentation | 140k Images                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| RedWeb-S           | Liu et al. (2021) | 2020 | In-the-wild | SOE                  | SCS         | Stereo Camera | Color                          | Salieny Mask                                                                  | 3179 Images                  |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| DUTLF-Depth        | Piao et al. (2019) | 2019 | In-the-wild | SOE                  | SCS         | Lytro Elum (Light Field (Stereo Camera) | Color, Depth | Salieny Mask                                                                  | 1200 Images                  |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Scene Flow         | Mayer et al. (2016) | 2016 | In-the-wild | SOE                  | -           | Synthetic   | Color                          | Optical Flow, Object Segmentation                                          | 2256 Scenes (39049 Frames)                                          |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| LFSD               | Frit et al. (2014) | 2015 | In-the-wild | SOE                  | SCS         | Lytro Elum (Light Field (Stereo Camera) | Color                          | Salieny Mask                                                                  | 100 Images                   |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| RGBD Salient Object Detection | Peng et al. (2014) | 2014 | In-the-wild | SOE                  | Structured Light | Kinect V1 | Color, Depth | Salieny Mask                                                                  | 1000 Images                   |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| MPT Sintel         | Butler et al. (2012) | 2012 | In-the-wild | SOE                  | -           | Synthetic   | Color, Depth                          | Optical Flow                                                                  | 25 Scenes (50 Frames/scene)                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| NYU+2-OC++          | Ramamonjisoa et al. (2020) | 2020 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth, Accelometer          | Occlusion Boundaries Maps, 2D Object Detection                               | 1449 Images From NYU/2                          |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Near-Collision Set | Mangik et al. (2019) | 2019 | Indoor     | SOE                  | SCS, LiDAR | LiDAR (N/A), Stereo Camera | Color, Depth | Salieny Mask                                                                  | 16368 Sequences                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| SfM-RGBD           | Camplani et al. (2017) | 2017 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth | Salieny Mask                                                                  | 33 sequences (15000 frames)                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| SUN RGB-D          | Song et al. (2015) | 2015 | Indoor     | SOE                  | Structured Light And TOF | Intel RealSense 3D Camera, Asus Xtion LIVE PRO, Kinect V1 and V2, ASUS Xtion ProLive RGB-D | Color, Depth | Semantic Segmentation, Object Detection And Pose | 10335 Images                                           |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| TUW                | Aldoma et al. (2014) | 2014 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth | Object Instance Recognition                         | 15 Sequences (163 Frames)                                         |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| Willow And Challenge | Aldoma et al. (2014) | 2014 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth | Object Instance Recognition                         | 24 Sequences (353 Frames) For Willow, 39 Sequences (176 Frames)            |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| An In Depth View of Salency | cpiladi et al. (2013) | 2013 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth | Salency Mask                                                                  | 80 Images                   |
|                    |                 |      |            |                      |             |             |                                        |                                                                              |                               |
| NYU Depth V2       | Silberman et al. (2012) | 2012 | Indoor     | SOE                  | Structured Light | Kinect V1 | Color, Depth, Accelerometer | Semantic Segmentation                                                      | 464 Scenes (407024 Frames) With 1449 Labeled Aligned RGB-D Images |

Continue on Next Page
parities. Despite using a custom structure light system, Middlebury 2014 contains improvements in the acquisition process.

4.1.4. Other

This sub-category contains all datasets that do not fit into the previous divisions. There is no sub-category in “Other” with more than four examples. Therefore, we did not create a specific sub-section for them.

All datasets here contain depth data and are divided into the following applications: novel view synthesis, foggy images for visibility restoration, relative depth between pairs of random points, object tracking, depth refinement for mirror surfaces, and synthesis of 4D RGB-D light field images. In Table 7, we display all these datasets and their respective application as a column of the table. The most cited dataset included here is:

**FRIDA2.** This dataset is a synthetic dataset of foggy images of the street view. It is formed by 330 synthetic images of 66 different scenes, where each image without fog is associated with four images that vary the intensity of the artificial fog presented in it (Tarel et al., 2012). Therefore, 66 images without fog have one depth map and four foggy images associated with it. FRIDA2 is a continuation of The Foggy Road Image Database (FRIDA) (Tarel et al., 2010), which has similar characteristics to FRIDA2, but fewer images (only 18 distinct scenes). These datasets are created for image enhancement in foggy images, trying to reduce the impact of the fog in the visibility of street scenes.

4.2. Body

In this category, all datasets are focused on body activities, such as action recognition, facial expression, hand activities, and sign language recognition. Here, we have only two sub-categories: the first one encompass full-body activities and the second one includes partial body parts, such as hands or face.

It is essential to notice that some of these datasets also include depth maps of the scene, but the focus of the dataset is on the Human Body (or part of it). Therefore, they are classified in this category.

4.2.1. Human Activities

This sub-category has all datasets focused on human activities, such as drinking, eating, playing tennis, and walking. Here, we have datasets that analyze actions for an individual (Wang et al., 2012b, 2014) or two-person interactions (Yun et al., 2012).

The majority of the works in the “Human Activities” sub-category are collected in controlled scenes, and we only found Hollywood 3D (Hadfield and Bowden, 2013) using in-the-wild datasets. The majority of the datasets are indoor scenes, but as they are centered on actions, they are classified in the “Scene Type” column as “Full Body”. The most common extra data is the person pose (or skeleton) of the people involved in the scene. Such information can help improve automatic action recognition algorithms. Datasets containing Human Activities are presented in Table 8. Next, we present three influential datasets in this sub-category:

**NTU RGB+D.** This dataset contains more than 50,000 video samples representing 60 distinct actions that are divided into three major groups: health-related actions (e.g., falling down, staggering), 40 daily actions (e.g., eating, drinking), 11 mutual actions (e.g., kicking, hugging) (Shahroudy et al., 2016). Forty subjects aged between 10 and 35 performed the actions in this dataset. The dataset was collected using three Kinect v2 from different horizontal views and is available with RGB, Depth, in-
| Dataset Name   | Ref.             | Year | Scene Type | Application | Sensor Name | Sensor Type | Data Modalities                                      | Extra Data                                      | Images/Scenes       |
|---------------|-----------------|------|------------|-------------|-------------|-------------|-----------------------------------------------------|------------------------------------------------|---------------------|
| –             | Wu et al. [2021]| 2020| Aerial     | SOR and SOE | -           | Synthetic   | Color, Depth                                       | Normal Maps, Edges, Semantic Labels             | 15 Scenes (144k Images) |
| EventScape    | Gebring et al.  | 2021| Driving    | SOR and SOE | -           | Synthetic   | Color, Depth                                       | Semantic Segmentation, Navigation Data (Position, Orientation, Angular Velocity, Etc.) | 758 Sequences       |
| KITTI-360     | Liao et al.     | 2021| Driving    | SOR and SOE | SCS, LiDAR  | Velodyne (LiDAR) Points Cloud, Stereo Camera      | Color, Depth, GPS, IMU                          | 2D-object Detection, 3D-object Detection, Tracking, Instance Segmentation, Optical Flow | These Are Not In Necessary In The Same Dataset | 11 Sequences To Over 320k Images And 100k Laser Scan |
| DDAD          | Ouazoulil et al. | 2020| Driving    | SOR and SOE | LiDAR       | Luminar-H2 LiDAR                                  | Color, Depth                                     | Instance Segmentation                              | 150 Scenes (12250 Frames) |
| Lyft Level 5  | Houston et al.  | 2020| Driving    | SOR and SOE | SCS, LiDAR  | 3 LiDAR (40 And 64-beam LiDAR), 5 Radars, Stereo Camera | Color, Depth, Radar                               | 3D Object Detection                                 | 170k Scenes (25 Seconds Each) |
| NuScenes      | Caesar et al.   | 2020| Driving    | SOR and SOE | LiDAR       | N/A                                                  | Color, Depth, Radar, IMU                         | 3D Object Detection, Semantic Segmentation        | 1000 Scenes (20 Seconds Each), 1.4M Images And 390k LiDAR Sweeps | |
| Woodscape     | Yingamani et al.| 2019| Driving    | SOR and SOE | LiDAR       | Velodyne HDL-64E                                   | Color, IMU, GPS, Depth                           | Instance Segmentation, 2D Object Detection         | Semantic Segmentation, Optical Flow                 | 50 Sequences (100k Frames) |
| Virtual Kitti | Gaidon et al.   | 2016| Driving    | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Semantic Segmentation, Instance Segmentation, Optical Flow | 50 Videos (21260 Frames) |
| KITTI         | Geiger et al.   | 2012| Driving    | SOR and SOE | SCS, LiDAR  | Velodyne (LiDAR) Points Cloud, Stereo Camera      | Color, Grayscale, Depth, GPS, IMU                 | Instance Segmentation                              | 61 Scenes (42746 Frames) |
| Hypersim      | Roberts et al.  | 2021| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Normal Maps, Instance Segmentation, Diffuse Reflectance | 461 Scenes (77400 Images) |
| RoboTHOR      | Deitske et al.  | 2020| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Instance Segmentation                              | 75 Scenes |
| Structured3D  | Zheng et al.    | 2020| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Object Detection, Semantic Segmentation            | 3500 Scenes With 21835 Rooms (196515 Frames) |
| Replica       | Straub et al.   | 2019| Indoor     | SOR and SOE | Structured Light | N/A                                                  | Color, Depth, IMU, Grayscale Camera            | Normal Maps, Semantic Segmentation                  | 18 Scenes |
| Gibson        | Xia et al.      | 2018| Indoor     | SOR and SOE | LiDAR, Structured Light                          | NavVis, Matterport Camera, DotProduct                 | Color, Depth                                           | Normal Maps, Semantic Segmentation                  | 572 Scenes. 1400 Floor Spaces From 572 Buildings |
| InterioNet    | Li et al.       | 2018| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth, IMU                          | Normal Maps, Semantic Segmentation                  | 20 Million Images |
| Taskonomy     | Garcia-Hernando et al. | 2018| Indoor     | SOR and SOE | Structured Light | N/A                                                  | Color, Depth                                           | 25 Tags (Normals Maps, Semantic Segmentation, Scene Classification, Etc.) | 4.5 Million Scenes |
| AVD           | Ammirato et al. | 2017| Indoor     | SOR and SOE | TOF         | Kinect V2                                           | Color, Depth                                           | Object Detection                                   | 15 Scenes (over 30k Images) |
| MatterPort3D  | Chang et al.    | 2017| Indoor     | SOR and SOE | SCS, Structured Light                           | Matterport Camera, Stereo Camera                       | Color, Depth                                           | Semantic Segmentation, 3D Semantic-voxel Segmentation | 90 Scenes, 10800 Panoramic Views (194400 Images) |
| ScanNet       | Dai et al.      | 2017| Indoor     | SOR and SOE | Structured Light                                    | Occupilal Structure Sensor - Similar to Kinect V1 | Color, Depth                                           | 3D Semantic-voxel Segmentation                     | 1513 Sequences (over 2.5 Million Frames) |
| SceneNet RGB-D| McCormac et al. | 2017| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Instance Segmentation, Optical Flow                 | 15K Trajectories (scenes) (15M Images) |
| SunCG         | Song et al.     | 2017| Indoor     | SOR and SOE | -           | Synthetic                                           | Color, Depth                                           | Semantic Segmentation                  | 45622 Scenes |
| GMU Kitchen   | Georgakis et al.| 2016| Indoor     | SOR and SOE | TOF         | Kinect V2                                           | Color, Depth                                           | Object Detection                                   | 9 Scenes (6735 Frames) |

Table 5: Datasets of “SLAM, Odometry, or Reconstruction” and “Segmentation or Other Extra Information” Categories
frared (IR) sequences, and person pose (skeleton) information. The authors extended the NTU RGB+D to a new dataset called NTU RGB+D 120, which contains other 60 classes and 57,600 samples, also containing the same capturing system and data modalities as the previous dataset [Liu et al., 2019].

**MSR DailyActivity3D Dataset.** This dataset covers sixteen different activities: drink, eat, read a book, call cellphone, write on a paper, use a laptop, use a vacuum cleaner, cheer up, sit still, toss paper, play games, lie down on a sofa, walk, play guitar, stand up, and sit down [Wang et al., 2012b]. Ten subjects performed each action twice: one for standing and one for sitting position. This dataset also includes person pose information for each frame. The authors used the Kinect v1 to acquire the depth of the scenes.

**MSR Action3D.** This dataset covers twenty different actions performed by ten subjects. Each action was performed two to three times, resulting in 557 filtered sequences and 23,797 frames [Li et al., 2010]. The actions are divided into three sets, where the first categorize actions with similar movements. The third set is composed by complex actions together. All sequences were acquired using Kinect v1 sensor.

### 4.2.2. Gestures (Partial Body)

Here, we grouped all works that involve human actions or activities and have data available for human body parts, such as arms, head, and hand. There is a wide variety of dataset purposes in this sub-category, such as action recognition based on a first-person view (no torso/head parts available in video) [Tang et al., 2017], salad preparation [Stein and McKenna, 2013], hand-pose information [Tompson et al., 2014], and sign language recognition [Wang et al., 2012a].

The most cited datasets in this sub-category include:

**NYU Hand Pose Dataset.** This dataset was captured using three Kinect v1, with two side views and a frontal view. The authors also re-created a synthetic hand pose for each view [Tompson et al., 2014], and made available 36 hand point locations for each frame. Three people acquired the data: one person used for training and the other two for testing, leading to over 80 thousand acquired frames.

**MSR Gesture3D.** This dataset contains sign language gestures. The authors collected 12 dynamic American Sign Language (ASL) gestures from ten people. The dataset was captured using Kinect v1, and has 336 sequences since each person performed multiple recordings of all selected signs. The authors performed a hand segmentation, and depth information is available only for the segmented hand regions. Background and body portions below the wrist were removed.

### 4.3. Medical

In this category, we present datasets that are from any part of the medical field. The exclusion criteria removed most of the datasets found here because these contained only private data. For instance, we collected eleven datasets containing endoscopic data, but only three meets all criteria to be included in our work. This situation is common in medical applications as sharing medical information requires regulated procedures.

We found only four datasets available in this category, of which three of them contain endoscopic data and one contains 3D models of the iris. The most cited dataset in containing depth information in the medical field is:

**Colonoscopy CG Dataset.** This dataset is composed of endoscopic data of the colon. To the best of our knowledge, this is the most frequent type of data that contains depth maps in the Medical category, even if analyzing datasets with non-shared data. The authors generated a synthetic dataset using Unity graphic engine based on a human CT colonography scan. They extracted a surface mesh using manual segmentation and meshing [Rau et al., 2019]. Their work also proposed and tested an algorithm in real data, but this data is not available for the community thus not included in this paper.
5. Discussion

The datasets presented in Section 4 constitute a collection of different scenes, sensors, and activities. We provide information about the Sensor Type, Number of Images/Scenes, Scene Type, Sensor Name, and Data Modalities available for each dataset. Unlike previous surveys of RGB-D datasets (Firman, 2016), we do not categorize the datasets regarding their realism since this is a subjective criterion and it is up to the researcher who will analyze the datasets to decide. Despite the variety of datasets presented, we identified common tendencies in all areas and discussed them in this section.

Although synthetic data is becoming more present each time, the usage of real data is presented in the majority of the datasets. Comparing the 2016-2018 to the 2019-2021 trienniums, we found a 50% increase in the numbers of datasets containing synthetic data. Synthetic datasets are usually cheaper to produce than performing real data acquisition because extra annotations, e.g., semantic segmentation or object tracking, are automatically generated. On the other hand, complex scene annotations for real data are costly, especially in scenes such as driving and aerial.

Synthetic datasets were initially created using simulators (Tarel et al., 2010, 2012), but these simulators were distinct to real-world scenarios since the computational power of the machines was limited. Hence, it was not possible to generate consistent and realistic datasets for complex scenes. Recently, realistic simulators were created for driving scenes, such as DrivingStereo - Driving Stereo Urban Virtual (UDV), which can produce consistent and realistic datasets for complex scenes. Additionally, realistic simulators were created for aerial scenarios, such as AirSim, which can produce consistent and realistic datasets for complex scenes.
as CARLA (Dosovitskiy et al., 2017), Nvidia Drive Sim\(^2\) and indoor scenes, such as Habitat (Szot et al., 2021; Savva et al., 2019). Despite the usage of simulators, other datasets rely on game engines or general computer graphics engines to build their systems, such as SYNTHIA (Ros et al., 2016). Virtual KITTI (Gaidon et al., 2016), and Virtual KITTI 2 (Cabon et al., 2020) that used Unity\(^3\) as graphic engine, and GTA-SfM (Wang et al., 2017) that uses scenes from the game GTAV.

The usage of synthetic data has been combined with real data to produce more complex scenes. These are applied especially for techniques that explore the generalization of their methods in non-expected scenes, i.e., using datasets not used in the training step (Ummenhofer et al., 2017; Ranftl et al., 2020; Eftekhar and Zisserman, 2021).

These papers combine datasets containing different types of acquisition and scenes to produce generalizable models. Ranftl et al. (2020) created multiple cross-dataset training strategies, and its combination of datasets with more images — called MIX5— contains data from DIML, MegaDepth, Red-Web, WSVD, and 3D Movies datasets. Ranftl et al. (2021b) expanded this combination, creating the MIX6 cross-dataset set containing about 1.4 million training images. Both works were evaluated using a mixture of testing datasets. The robustness of the models are also evaluated in a cross-dataset strategy for estimating depth from a monocular video (Kopf et al., 2021), and instead of testing in multiple types of scenes, Ji et al. (2021) combined distinct datasets of the same type of scene to improve the results for the indoor environment.

Recently, domain adaptation has been applied to improve the performance of the combination of datasets in the training step (Guo et al., 2018; Atapour-Abarghouei and Breckon, 2018; Zhao et al., 2019). Atapour-Abarghouei and Breckon (2018), for instance, combines one synthetic and one real dataset using domain adaptation to improve the result of training. They claim that directly using synthetic data may not improve the results for realistic data evaluation due to dataset bias. They adapt the domain of a synthetic dataset to a real dataset using Style Transfer and combine them to train their models. Zhao et al. (2019) also performs domain adaptation, and they claim that due to the lack of paired synthetic and real images, the synthetic-to-realistic image translation adds distortions to the depth estimation. They overcome this difficulty by exploring a more complex training procedure involving synthetic-to-realistic and realistic-to-synthetic translations. To generate more realistic synthetic data, Su et al. (2015) proposed the use of 3D CAD Models to produce 2D synthetic images, since these CAD Models allow multiple viewpoints and complete control of the deformations in the modeled objects to increase the variability of the created dataset. Planche et al. (2017) also used 3D CAD Models, but they intended to create realistic depth data from the 3D objects. They proposed a framework that simulates real distortion factors of depth data acquisition, e.g., material reflectance and sensor noise, to generate reliable depth data. In addition to using synthetic data, domain adaptation could also be applied to real-to-real translation (Lopez-Rodriguez and Mikolajczyk, 2020; Hornauer et al., 2021) since the dataset bias also affects distinct real datasets, especially by variations of scale and capture’s position of the scenes (Torralba and Efros, 2011).

6. Conclusions

In this work, we presented a survey of publicly available image datasets that contain depth information. We categorized and summarized over 200 datasets based on the image scenes, sensors used to collect the depth information, and the different applications for which these datasets can be used. Almost half of the datasets we describe were proposed after the publication of the last survey (Cai et al., 2017). The new datasets expand the scope of applications that depth datasets can be used for, such as medical applications. The new datasets also expand the quality and quantity of data for other areas.
| Dataset Name | Ref. | Year | Scene Type | Sensor Type | Sensor Name | Data Modalities | Extra Data | Images/Scenes |
|--------------|------|------|------------|-------------|-------------|----------------|------------|---------------|
| Depth 2 Height | Yin and Zhou [2020] | 2020 | Full Body | TOF | Kinect V2 | Color, Depth | - | 2136 Images |
| SOR3D-AFF | Thermos et al. [2020] | 2020 | Full Body | TOF | Kinect V2 | Color, Depth | - | 1201 Sequences |
| NTU RGB+D 120 | Lu et al. [2019] | 2019 | Full Body | TOF | Kinect V2 | Color, Depth, IR | Person Pose (Skeleton) | 111480 Sequences |
| -- | Tang et al. [2019] | 2019 | Full Body | TOF | Kinect V2 | Color, Depth | - | 800 Frames For Each Person (26 People) |
| CMDFALL | Tran et al. [2018] | 2018 | Full Body | Structured Light | Kinect V1 | Color, Depth, Accelerometer | - | 20 Sequences |
| UESTC | Ji et al. [2019] | 2018 | Full Body | TOF | Kinect V2 | Color, Depth | - | 256600 Sequences |
| UOW Online Action3D | Tang et al. [2018] | 2018 | Full Body | TOF | Kinect V2 | Color, Depth | Person Pose (Skeleton) | 20 Sequences (20 Participants Performing Multiple Actions In A Sequence) |
| PKU-MMD | Chunhui et al. [2017] | 2017 | Full Body | TOF | Kinect V2 | Color, Depth, Accelerometer | Person Pose (Skeleton) | 3076 Sequences |
| TVPR | Liscotti et al. [2017] | 2017 | Full Body | Structured Light | Asus Xtion Pro Live | Color, Depth | - | 23 Sequences (100 People, 2004 Secs) |
| Chalaeum LAP IsoGD | Wan et al. [2016] | 2016 | Full Body | Structured Light | Kinect V1 | Color, Depth | - | 47953 Sequences |
| G3D | Bloom et al. [2016] | 2016 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) Semantic Segmentation | 7 Sequences (Multiple Actions Per Sequence) |
| HHOI | Shu et al. [2016] | 2016 | Full Body | TOF | Kinect V2 | Color, Depth | Person Pose (Skeleton) | 8 Actors Recorded Interactions, Each Interaction Lasts 2-7 Seconds Presented At 10-15 Fps |
| ISR-UoL 3D Social Activity | Coppola et al. [2016] | 2016 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 10 Sequences |
| NTU RGB+D | Shahroudy et al. [2015a] | 2015 | Full Body | TOF | Kinect V2 | Color, Depth, IR | Person Pose (Skeleton) | 56880 Sequences |
| TST Fall Detection V2 | Gasparrin et al. [2015b] | 2015 | Full Body | TOF | Kinect V2 | Color, Depth, Accelerometer | Person Pose (Skeleton) | 264 Scenes |
| UOW LargeScale Combined Action3D | Zhang et al. [2016a] | 2016 | Full Body | TOF | Kinect V2 | Color, Depth | Person Pose (Skeleton) | 4953 Sequences |
| CMU Panoptic | Joo et al. [2017] | 2015 | Full Body | TOF | Kinect V2 | Color, Depth | 3D Skeleton | 65 Sequences (5.5 Capture Hours) |
| SYSU 3D HOI | He et al. [2017] | 2015 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 480 Sequences |
| TST Intake Monitoring V1 | Gasparrin et al. [2015a] | 2015 | Full Body | Structured Light | Kinect V1 | Color, Depth | - | 48 Sequences |
| TST Intake Monitoring V2 | Gasparrin et al. [2015a] | 2015 | Full Body | Structured Light | Kinect V1 | Color, Depth | - | 60 Sequences |
| TST TUG DataBase | Cipipitieli et al. [2015] | 2015 | Full Body | TOF | Kinect V2 | Color, Depth, Accelerometer | Person Pose (Skeleton) | 60 Sequences |
| UTD-MHAD | Chen et al. [2015] | 2015 | Full Body | Structured Light | Kinect V1 | Color, Depth, Accelerometer, Motion Capture (ms) Camera | Person Pose (Skeleton) | 861 Sequences |
| Human3.6M | Ionescu et al. [2014] | 2014 | Full Body | TOF | MESA Imaging SR4000 from SwissRanger | Color, Depth, Accelerometer, Motion Capture (ms) Camera | Person Pose (Skeleton) | 447260 RGB-D Frames (almost 3.6M RGB Frames) |
| KARD | Gaglione et al. [2015] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 540 Sequences |
| LIRIS | Wolf et al. [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth | - | 180 Sequences |
| MAD | Huang et al. [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 40 Sequences |
| Northwestern-UCLA Multiview Action 3D | Wang et al. [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth | - | 1473 Sequences |
| Online RAGB Action Dataset (ORAGBD) | Yu et al. [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 48 Sequences |
| TST Fall Detection V1 | Gasparrin et al. [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth, Accelerometer | Person Pose (Skeleton) | 20 Sequences |
| UR Fall Detection | Kwolek and Kempka [2014] | 2014 | Full Body | Structured Light | Kinect V1 | Color, Depth, Accelerometer | - | 70 Sequences |
| Chalearum Multimodal Gesture Recognition | Escalera et al. [2013] | 2013 | Full Body | Structured Light | Kinect V1 | Color, Depth, Audio | User Mask, Person Pose (Skeleton) | 707 Sequences (1720800 Frames) |
| Florence 3D Actions | Seidenari et al. [2013] | 2013 | Full Body | Structured Light | Kinect V1 | Color, Depth | Person Pose (Skeleton) | 215 Sequences |

Continue on Next Page
We also presented different forms of acquiring depth information from a scene. We expect that this explanation could be used in conjunction with extra information of the datasets to allow researchers to choose the ones that best fulfill their needs. Researchers of zero-shot learning trying to increase generalization capabilities for their model could also benefit from our work since they may select distinct datasets in terms of sensor type, application, and scene type for training and evaluating their methods.

CRediT authorship contribution statement

Alexandre Lopes: Conceptualization, Formal analysis, Investigation, Methodology, Writing - review & editing. Roberto Souza: Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing. Helio Pedrin: Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful to the National Council for Scientific and Technological Development, Brazil (CNPq grant 309330/2018-1). Roberto Souza thanks the Natural Sciences and Engineering Research Council (NSERC - RGPIN-2021-02867) for ongoing operational support.

References

Aksoy, E.E., Tamosiunaite, M., Wörgötter, F., 2015. Model-Free Incremental Learning Of The Semantics Of Manipulation Actions. Robotics and Autonomous Systems, 118–133.

Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzent senko, V., Alvarez, F., Zarpalas, D., Daras, P., 2021. Pano3D: A Holistic Benchmark And A Solid Baseline for 360Deg Depth Estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3727–3737.

Almada, A., Füllhammer, T., Vincze, M., 2014. Automation Of “Ground Truth” Annotation for Multi-View RGB-D Object Instance Recognition Datasets, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5016–5023.

Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zelek, T., Xia, W., 2021. Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. arXiv preprint arXiv:2101.01133 , 1–17.

Ammirato, P., Poisron, P., Park, E., Kosecka, J., Berg, A.C., 2017. A Dataset for Developing and Benchmarking Active Vision, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 1378–1385.

Antequera, M.L., Gargallo, H., Hofinger, M., Buló, S.R., Kuang, Y., Kontschieder, P., 2020. Mapillary Planet-Scale Depth Dataset, in: European Conference on Computer Vision (ECCV), pp. 589–604.

Armeni, I., Sax, S., Zami, A.R., Savaresi, S., 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. arXiv preprint arXiv:1702.01105 , 1–9.

Atapour-Abbarghouei, A., Breckon, T.P., 2018. Real-time Monocular Depth Estimation using synthetic Data with Domain Adaptation via Image Style Transfer, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2800–2810.

Bagdanov, A.D., Del Bimbo, A., Masi, I., 2011. The Florence 2D/3D Face Dataset, in: Joint ACM Workshop on Human Gesture and Behavior Understanding, p. 79–80.

Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V., 2012. Re-Identification With RGB-D Sensors, in: European Conference on Computer Vision (ECCV), pp. 433–442.

References

Aksoy, E.E., Tamosiunaite, M., Wörgötter, F., 2015. Model-Free Incremental Learning Of The Semantics Of Manipulation Actions. Robotics and Autonomous Systems, 118–133.

Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzent sensko, V., Alvarez, F., Zarpalas, D., Daras, P., 2021. Pano3D: A Holistic Benchmark And A Solid Baseline for 360Deg Depth Estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3727–3737.

Almada, A., Füllhammer, T., Vincze, M., 2014. Automation Of “Ground Truth” Annotation for Multi-View RGB-D Object Instance Recognition Datasets, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5016–5023.

Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zelek, T., Xia, W., 2021. Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. arXiv preprint arXiv:2101.01133 , 1–17.

Ammirato, P., Poisron, P., Park, E., Kosecka, J., Berg, A.C., 2017. A Dataset for Developing and Benchmarking Active Vision, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 1378–1385.

Antequera, M.L., Gargallo, H., Hofinger, M., Buló, S.R., Kuang, Y., Kontschieder, P., 2020. Mapillary Planet-Scale Depth Dataset, in: European Conference on Computer Vision (ECCV), pp. 589–604.

Armeni, I., Sax, S., Zami, A.R., Savaresi, S., 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. arXiv preprint arXiv:1702.01105 , 1–9.

Atapour-Abbarghouei, A., Breckon, T.P., 2018. Real-time Monocular Depth Estimation using synthetic Data with Domain Adaptation via Image Style Transfer, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2800–2810.

Bagdanov, A.D., Del Bimbo, A., Masi, I., 2011. The Florence 2D/3D Face Dataset, in: Joint ACM Workshop on Human Gesture and Behavior Understanding, p. 79–80.

Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V., 2012. Re-Identification With RGB-D Sensors, in: European Conference on Computer Vision (ECCV), pp. 433–442.

References

Aksoy, E.E., Tamosiunaite, M., Wörgötter, F., 2015. Model-Free Incremental Learning Of The Semantics Of Manipulation Actions. Robotics and Autonomous Systems, 118–133.

Albanis, G., Zioulis, N., Drakoulis, P., Gkitsas, V., Sterzent senko, V., Alvarez, F., Zarpalas, D., Daras, P., 2021. Pano3D: A Holistic Benchmark And A Solid Baseline for 360Deg Depth Estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3727–3737.

Almada, A., Füllhammer, T., Vincze, M., 2014. Automation Of “Ground Truth” Annotation for Multi-View RGB-D Object Instance Recognition Datasets, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5016–5023.

Allan, M., Mcleod, J., Wang, C., Rosenthal, J.C., Hu, Z., Gard, N., Eisert, P., Fu, K.X., Zelek, T., Xia, W., 2021. Stereo Correspondence and Reconstruction of Endoscopic Data Challenge. arXiv preprint arXiv:2101.01133 , 1–17.

Ammirato, P., Poisron, P., Park, E., Kosecka, J., Berg, A.C., 2017. A Dataset for Developing and Benchmarking Active Vision, in: IEEE International Conference on Robotics and Automation (ICRA), pp. 1378–1385.

Antequera, M.L., Gargallo, H., Hofinger, M., Buló, S.R., Kuang, Y., Kontschieder, P., 2020. Mapillary Planet-Scale Depth Dataset, in: European Conference on Computer Vision (ECCV), pp. 589–604.

Armeni, I., Sax, S., Zami, A.R., Savaresi, S., 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. arXiv preprint arXiv:1702.01105 , 1–9.

Atapour-Abbarghouei, A., Breckon, T.P., 2018. Real-time Monocular Depth Estimation using synthetic Data with Domain Adaptation via Image Style Transfer, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2800–2810.
| Dataset Name            | Ref.          | Year | Scene Type          | Sensor Type       | Sensor Name                  | Data Modalities                       | Extra Data                  | Images/Scenes |
|------------------------|---------------|------|---------------------|-------------------|------------------------------|---------------------------------------|----------------------------|----------------|
| Bimanual Actions       | Dreher et al. | 2020 | Part Of Body (hand, Head, Etc.) | Structured Light  | IntelRealSense SR300/Synthetic | Color, Depth                         | -                          | 16016 Images   |
| RGB2Hands              | Wang et al.   | 2020 | Part Of Body (hand, Head, Etc.) | Structured Light  | IntelRealSense SR300/Synthetic | Color, Depth                         | -                          | 25 Scenes      |
| ObMan                  | Hasson et al. | 2019 | Part Of Body (hand, Head, Etc.) | -                 | Synthetic                    | Color, Depth                         | 3D Hand Keypoints, Object Segmentation, Hand Segmentation | 150k Images     |
| EgoGesture             | Zhang et al.  | 2018 | Part Of Body (hand, Head, Etc.) | Structured Light  | IntelRealSense SR300/Synthetic | Color, Depth                         | -                          | 2081 Sequences |
|                        | Garcia-Hernando et al. | 2018 | Part Of Body (hand, Head, Etc.) | Structured Light  | IntelRealSense SR300/Synthetic | Color, Depth                         | -                          | 1175 Sequences |
| BigHand2.2M            | Wun et al.    | 2017 | Part Of Body (hand, Head, Etc.) | Structured Light  | IntelRealSense SR300          | Color, Depth, 6D Magnetic Sensor      | -                          | N/A Sequences  |
| Pandora                 | Bonghi et al. | 2017 | Part Of Body (hand, Head, Etc.) | TOF               | Kinect V2                    | Color, Depth                         | Upper Body Part Pose (Skeleton) | 100 Sequences  |
| RHD                    | Zimmermann and Bix 2017 | 2017 | Part Of Body (hand, Head, Etc.) | -                 | Synthetic                    | Color, Depth                         | Segmentation, Keypoints      | 43986 Images   |
| THU-READ               | Tang et al.   | 2017 | Part Of Body (hand, Head, Etc.) | Structured Light  | PrimeSense Carmine           | Color, Depth                         | -                          | 1920 Sequences |
| STB                    | Zhang et al.  | 2017 | Part Of Body (hand, Head, Etc.) | Structured Light  | Intel Real Sense F200, Stereo Camera | Color, Depth                         | -                          | 12 Sequences   |
| Creative Sen3D         | Memo et al.   | 2015 | Part Of Body (hand, Head, Etc.) | Structured Light  | Creative Sen3D               | Color, Depth                         | -                          | 1320 Sequences |
| EYEDIAP                | Funes Mora et al. | 2014 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                         | Eye Points, Head Pose        | 94 Sequences   |
| Eurecom Kinect Face    | Man et al.    | 2014 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                         | Face Points                  | 936 Sequences  |
| Hand Gesture           | Marin et al.  | 2014 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                         | -                          | 1400 Sequences |
| MANIC                  | Akeley et al. | 2015 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                         | -                          | 103 Sequences  |
| NYU Hand Pose          | Timpson et al. | 2014 | Part Of Body (hand, Head, Etc.) | Structured Light  | Structured Light             | Color, Depth                        | Hand Pose                   | 81009 Frames   |
| 3D MAD                 | Nesi and Marcel | 2013 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                        | Eye Points                   | 255 Sequences  |
| 50 Salads              | Stein and McKenz | 2013 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth, Accelerometer          | Activity Classification     | 50 Sequences   |
| Dexter 1               | Strahl et al. | 2013 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1, Creative Gesture Camera, Stereo Camera | Color, Depth                        | -                          | 7 Sequences    |
|                        | Xu and Cheng  | 2013 | Part Of Body (hand, Head, Etc.) | TOF               | SoftKinetic DS 325           | Color, Depth, Measured Hand          | -                          | 870 Images (30 Subjects) |
| MSRC-12                | Fisher-gill et al. | 2012 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                        | -                          | 594 Sequences  |
| Kinect Gesture         | Fothergill et al. | 2012 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                        | -                          | 719359 Frames  |
| MSR Gesture3D          | Wang et al.   | 2012 | Part Of Body (hand, Head, Etc.) | Structured Light  | Kinect V1                    | Color, Depth                        | -                          | 336 Sequences  |
| Florence 3D Faces      | Bagdanov et al. | 2011 | Part Of Body (hand, Head, Etc.) | -                 | Synthetic                    | Color, Depth                        | -                          | 53 People (N/A Frames/Seq) |

Table 9: Datasets of “Gestures (Partial Body) sub-category”

Table 10: Datasets of “Medical” Category.

Benalcazar, D.P., Zambrano, J.E., Bastias, D., Perez, C.A., Bowyer, K.W., 2020. A 3D Iris Scanner From A Single Image Using Convolutional Neural Networks. IEEE Access , 98584–98599.

Berman, D., Levy, D., Avidan, S., Treibitz, T., 2021. Underwater Single Image Color Restoration Using Haze-Lines And A New Quantitative Dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2822–2837.

Bianco-Claraco, J.L., Moreno-Duenas, F.A., González-Jiménez, J., 2014. The Málaga Urban Dataset: High-Rate Stereo and LiDAR In A Realistic Urban Scenario. The International Journal of Robotics Research , 207–214.

Bloom, V., Argyriou, V., Makris, D., 2016. Hierarchical Transfer Learning for Online Recognition of Compound Actions. Computer Vision and Image Understanding, 62–72.

Borghi, G., Venturelli, M., Vezzani, R., Cucchiara, R., 2017. Poseidon: Face-From-Depth for Driver Pose Estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4661–4670.

Borrás, R., Lapedriza, A., Ignac, L., 2012. Depth Information In Human Gait Understanding, 62–72.

Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R., 2016. The Euroc Micro Aerial Vehicle Datasets. The
