A Review of Benchmark Datasets and Training Loss Functions in Neural Depth Estimation

FAISAL KHAN¹, SHAHID HUSSAIN², SHUBHAJIT BASAK³, MOHAMED MOUSTAFA¹, AND PETER CORCORAN¹ (Fellow, IEEE)

¹Department of Electronic Engineering, College of Science and Engineering, National University of Ireland Galway, Galway, H91 TK33 Ireland
²Data Science Institute, National University of Ireland Galway, Galway H91 TK33, Ireland
³School of Computer Science, National University of Ireland Galway, Galway H91 TK33, Ireland

Corresponding author: Faisal Khan (f.khan4@nuigalway.ie).

This work was supported by the College of Science and Engineering, National University of Ireland Galway, Galway, H91TK33 Ireland; the Xperi Galway Block 5 Parkmore East Business Park, Galway, H91V0TX, Ireland; and the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

ABSTRACT In many applications, such as robotic perception, scene understanding, augmented reality, 3D reconstruction, and medical image analysis, depth from images is a fundamentally ill-posed problem. The success of depth estimation models relies on assembling a suitably large and diverse training dataset and on the selection of appropriate loss functions. It is critical for researchers in this field to be made aware of the wide range of publicly available depth datasets along with the properties of various loss functions that have been applied to depth estimation. Selection of the right training data combined with appropriate loss functions will accelerate new research and enable better comparison with state-of-the-art. Accordingly, this work offers a comprehensive review of available depth datasets as well as the loss functions that are applied in this problem domain. These depth datasets are categorised into five primary categories based on their application, namely (i) people detection and action recognition, (ii) faces and facial pose, (iii) perception-based navigation (i.e., street signs, roads), (iv) object and scene recognition, and (v) medical applications. The important characteristics and properties of each depth dataset are described and compared. A mixing strategy for depth datasets is presented in order to generalise model results across different environments and use cases. Furthermore, depth estimation loss functions that can help with training deep learning depth estimation models across different datasets are discussed. State-of-the-art deep learning-based depth estimation methods evaluations are presented for three of the most popular datasets. Finally, a discussion about challenges and future research along with recommendations for building comprehensive depth datasets will be presented as to help researchers in the selection of appropriate datasets and loss functions for evaluating their results and algorithms.

INDEX TERMS Depth Datasets, Depth Loss Function, Deep Learning, Depth Estimation.

I. INTRODUCTION

Depth estimation, the process of preserving 3D information of a scene using 2D information acquired by camera, can proof beneficial for many challenging computer-vision applications. Examples include human-machine interaction, robotics, augmented reality, object detection, pose estimation, semantic segmentation, and 3D reconstruction. Having access to ground truth depth information is valuable for developing robust guidance systems in autonomous vehicles, environment reconstruction, security, and image understanding where it is desirable to determine the primary objects and region with the imaged scene.

To this end, various methods have been developed to capture depth measurements as well as to research depth estimation using monocular or multi-view solutions, which aim to find the distance between scene objects and camera from a single or multiple point(s) of view relying on one or more images.
A. Application Classes of Depth Dataset

Datasets play a crucial role in scientific research, specifically for artificial intelligence models, datasets are the building block for analysing the performance and validating their results. Different datasets contain data captured in different environments (e.g., indoor vs outdoor scenes), of different objects, depth annotation types (relative, absolute, dense, sparse), accuracies (laser stereo, time-of-flight, synthetic data, structure-from-motion, human annotation), image quality, size, and camera settings. Every dataset has its own features and related problems and biases [1]. Large dataset collections from internet sources have many issues including quality of images, accuracy, and unknown camera parameters [2], [3]. High quality datasets can play an important role at enabling researchers to develop depth solutions for specific computer vision depth problems [4], [5].

Depth datasets are classified into various categories depending on particular task-based applications (i.e., indoor/outdoor, portrait/driver, half/full body scene, indoor small room, large street scene, large indoor scene, landscape/cityscape, and medical).

Structured light cameras, which give dense depth maps up to 10 meters, are commonly used to collect indoor depth information. They work by projecting a sequence of known patterns onto an object, and the deformation resulting from the object’s shape is then observed through a camera from some other direction. Depth information can then be extracted from the observed distortion’s disparity from the original projected pattern. The original Kinect sensor, also called Kinect v1, along with the Asus Xtion Pro, utilize this approach for depth capture [6]. Another commonly used technique is time-of-flight cameras, such as the Kinect v2, which relies on measuring the round-trip time for an emitted light using a sensor array and illumination unit [6]. Indoor places include locations such as offices, labs, corridors, study rooms, laboratories, and kitchens. Visual localization allows for intriguing applications like robot navigation and augmented reality by estimating the precise location of a camera. This is particularly useful in indoor environments where other localization technologies, such as Global Navigation Satellite System (GNSS), fail. Indoor spaces impose interesting tasks on visual localization methods (i.e., texture-less surfaces, occlusions due to people, large view-point changes, repetitive textures, and low light).

Outdoor depth datasets are typically collected with a specific application in mind such as autonomous vehicles and generally captured with customized sensor arrays consisting of multi or monocular cameras and Light Detection and Ranging (LiDAR) scanners. Outdoor place categories include street signs, forests, indoor/outdoor parking lots, urban areas, roads, residential areas, and coast areas. The primary applications of outdoor depth datasets involve perception tasks in the context of autonomous vehicles, semantic scene understanding, and 3D reconstruction.

Human faces are one of the most prevalent features in images, and thus are a key part of a lot of computer vision tasks. It is widely known in human skeletal anatomy that the eye-separation in a human face fall within a small range, thus given information of a camera’s field-of-view, it is feasible to calculate the distance-to-camera of a human subject with reasonable accuracy [7]. Human facial depth datasets include facial images, depth maps, images of the visible light spectrum (i.e., RGB), 3D depth maps, and head pose information. Deep neural networks can be trained to detect age, face, and gender using facial depth datasets, or to pick the optimum type of image for a specific task, such as facial recognition. It is also feasible to utilize data from people in random and frontal orientations to see if a facial recognition system can recognize faces from different perspectives [7], [8]. The face recognition system is typically divided into two different tasks in the computer vision field such face identification and face verification. The former is based on a one-to-many comparison to recognize the best match between a given face and a set of possibilities. While the latter uses a one-to-one comparison and can find whether the input item is of the same person’s face or not.

Depth datasets created for a medical application consist of multi-view frames, video, RGB, depth maps, calibration parameters, 2D and/or 3D pose annotations, and human bounding boxes. The data generated during surgeries can be used for medical image analysis and machine learning to observe, analyze, model and support staff activities and clinician in the operating rooms.

Ideally researchers should combine multiple datasets during training, validation, and testing to improve generalization, but care is needed when combining datasets with differing characteristics. The design and building blocks of the network are important, but the performance of the network is mostly determined by how it is trained which requires a diverse dataset and a suitable loss function.

B. Loss Functions for Depth Datasets

Another way to improve the deep network’s training results is by introducing an appropriate loss function. The loss function calculates the network output’s variance from the estimated output which is used to adjust the parameters of the deep network. This is achieved by backpropagating the error calculated using the loss function to the first layer in the training process, changing the network’s weights at each iteration.

A deep network must have a loss function. The loss function must be differentiable because of the back-propagation stage used in deep learning systems, which relies on propagating the gradients of the model’s error from the output layer back towards the first layer. An in-depth study of various loss functions for depth regression is proposed that can be used for both short and long-range depth datasets.
C. Research Contributions

This review aims to collect the available depth image datasets using bibliometric research by providing detailed information on the available datasets. Additionally, an easy and brief description is presented for each of the datasets to provide a basis for predicting depth estimation trends and explores their sub-areas; dataset popularity helps in identifying study areas that receive less attention.

The main scope of this study is to make it easier to navigate among the depth datasets and common loss functions that are frequently used in the depth estimation research. A list of popular datasets is compiled by looking through the publications indexed by the web of science library and IEEE Explore, as well as doing searches utilizing online search engines. These datasets are classified into different use case categories and present their detailed description such as (camera tracking, scene reconstruction, tracking, semantic, pose, video and recognition, streets, people i.e., identity recognition/faces, medical depth-based applications, indoor and outdoor scenes). The most popular datasets are highlighted, together with bibliographic information (such as the number of citations). Furthermore, different aspects of the datasets are compared, common characteristics of popular datasets are described, and key recommendations for generating depth estimation datasets are suggested. The dataset description, metadata, ground truth, and relevant information i.e. (year of publication, ground truth information, size of the images, type, objects per image and number of images) are all listed in a structured way for each dataset. Also, each loss function is described in a way that can help the research community choose a right loss function for their specific tasks.

The rest of the survey paper is organized as follows: Section 2 describes related work, primarily other studies or surveys in the field of depth estimation. The findings of a bibliometric study are provided in section 3. A comprehensive review of depth datasets is presented in Section 4. Section 5 describes common characteristics of popular datasets. Top five state-of-the-art (SoA) depth estimation methods on three most popular datasets are presented in Section 6. In section 7, popular depth estimation loss functions are studied. A brief overview, relevant research, problems, and future research prospects are presented in Section 8. A summary of the current review is offered in section 9, while sections 10 and 11 make broad recommendations for creating new datasets to achieve scientific importance and conclusion.

II. RELATED WORKS

In this section, a review of the current SoA research is provided for depth datasets. Next, an overview of available related depth estimation research and 3D reconstruction articles is presented, followed by depth from 2D, monocular, and depth from Stereo & Multi-View depth datasets.

A. Depth Datasets

The authors in [8] presented a detailed analysis of image-based depth estimation and 3D reconstruction. They provided details of existing systems, shortcomings, and reconstruction approaches while briefly introducing five publicly available datasets for depth estimation. However, due to several limitations, particularly hardware (e.g., sensors and optics limitations), the applicability of such datasets is questionable for future research. The authors in [9] looked at image segmentation research using deep learning with details of five public depth datasets and briefly discussed other segmentation datasets. The authors also point out sensor limitations and future research directions, but they don’t explain all the relevant datasets.

While the authors in [10] presented an analysis of a method that combines ten datasets for monocular depth estimation with results on ten datasets, a description for utilizing the datasets, however, is not presented. An overview of deep-learning algorithms for monocular depth estimation using two public datasets was published in [11]; they present the significance of using NYU-v2 and KITTI datasets and argue that comprehensive testing with other datasets is required.

Three types of depth estimation datasets were chosen and described in [12] for understanding depth estimation models. The application of deep learning algorithms with four primary depth datasets for monocular depth estimation was studied in [13]. However, some of the relevant datasets which may influence the performance were not given much importance. The authors in [14] surveyed deep learning-based monocular depth estimation algorithms in the visible spectrum by describing a total of seven visible spectrum datasets. Some of the existing review articles [15]–[20] focusing on depth estimation either from single or multiple views, but the accessibility of those datasets is unclear.

B. Depth estimation research and 3D reconstruction

One of the most useful intermediate representations for action in physical environments is depth information, however, activity depth estimation remains a challenging problem in computer vision. To solve it, one must exploit many, sometimes, visual cues, subtle, short-range or long-range context, along with their corresponding information. This calls for learning-based methods. Depth estimation methods have been shown in the SoA to be a potential solution to several of problems [10], [11], [15]. Accurate depth estimation approaches can help with understanding 3D scene geometry and 3D reconstruction, which is especially significant in cost-sensitive applications and use case applications [16]. A comprehensive review of 3D reconstruction research is proposed in [8], which focuses on the work that uses deep neural network-based methods to estimate the 3D shape either from single or multi-view images [21].
C. Depth from 2D, monocular Images

Estimating depth information from 2D images is one of the most important problems in the field of computer vision and image processing. Depth information can be applied in 2D to 3D reconstruction, scene refocusing, scene understanding, depth-based image editing, and 3D scene conversion. The problem of monocular depth estimation is currently best tackled with convolutional neural networks due to their properties that can be used particularly in cost-sensitive applications [22]. SoA monocular depth methods have been reviewed in [11], [17], [18], [23]-[25], which focus on both non-deep learning and deep learning methods.

D. Depth from Stereo & Multi-View

Depth from stereo or multi-view can be obtained by using two or more cameras. The main idea is that triangulation and stereo matching can be used to estimate the depth, which can be utilized in various tasks such as robotic navigation, different object grasp, collision avoidance, or broadcasting and multimedia. Various methods have been studied in [2], [4], [8], [20], [26] that focus on depth estimation from both stereo and multi-view images.

III. METHODOLOGY FOR REVISING DEPTH DATASETS AND LOSS FUNCTIONS EMPLOYED IN LITERATURE

Utilizing the most suitable dataset for a given task is a basic assumption for the effective training and validation of any scientific method. In the domain of depth estimation research, the lack of publicly available depth estimation datasets and loss functions present challenges for researchers for their specific task or use-case.

This section aims to provide an in-depth explanation of the methodology used to search for and collect more than 40 popular datasets and loss functions which is presented in this review. The authors defined popularity based on the citation rank within the research areas and provide a detailed list of collected datasets and loss functions, as well as reviewed papers, in subsequent sections.

A. EXPLORING THE IMAGE DEPTH RELATED RESEARCH

There are numerous literature sources related to depth estimation. This study focuses on research publications that involve depth estimation tasks such as smart mobility-based road navigation, object detection, 3D reconstruction, robotics, and self-driving cars. The search methodology illustrated in Fig. 1 is adopted as to concentrate on the most relevant papers as well as leverage popular libraries and search tools such as Web of Science, Google Scholar, and IEEE Engineering online libraries.

Keywords such as "depth estimation and 3D reconstruction", "depth datasets, databases", "monocular and multi view depth estimation methods" were used as search criteria which helped in identifying 634 relevant journal papers. The selection of papers was based on three main factors: (i) Computer vision, engineering, deep learning, imaging technology, autonomous vehicles and robotics, 3D reconstruction, (ii) Science citation index, and (iii) English language.

B. PRIMARY STUDIES AND ASSESSMENT OF RESEARCH QUALITY

Following the research methodology (Fig. 1), the initial filter search using the datasets keyword retrieved 321 results for depth datasets and 212 results for loss functions out of 634 papers, the results were further analysed by title and abstract which filtered out 145 and 104 research articles respectively. Next, it is analysed that the text with the criteria being the selection of those articles in which the authors discussed at least one depth image datasets and loss function, carried out manually by reading the selected research articles. Such analysis helped in further reducing the number of papers to 92 and 80, which were further filtered down to the most relevant 52 and 48 articles using full-text-based selection criteria. As per the last stage’s criteria, the following categories of articles are excluded:

1. Those publications that are not directly related to depth estimation research. Examples include studies on 3D reconstruction or segmentation tasks datasets.
2. Reproductions or the same research work appearing in several places.
3. Studies that are concerned with human depth but do not make use of any depth datasets (e.g., review studies).

C. ANALYSIS OF THE MOST RELEVANT DATASETS

The methodology discovered that about 61% of the total papers in this domain considered at least one dataset in their experimental study. Additionally, 51% of the publications considered two or more than two datasets. Fig. 2 shows the results, where it is highlighted that the overall number of citations for the most popular datasets. The figure indicates that the most highly ranked depth datasets are KITTI, Cityscapes, and NYU-V2, with a citation count of 141, 94, and 78 in 120, 70, and 52 papers, respectively. This implies that about 25% of the studies considered these datasets for depth estimation tasks. These datasets are considered benchmark datasets in about 242 (77%) research studies.

The descriptions and comparisons of numerous criteria used to assist in navigating current publicly available datasets are presented by focusing on the usefulness of the datasets for specific study areas. The nature of the data imposes several restrictions on the availability of the datasets to the public. To assess the current availability of each dataset, their accessibility, in terms of access and obtaining a copy, is confirmed manually by the authors for each dataset. The test for access to each of the datasets included checking free access and an email-based inquiry to the host institution.
FIGURE 1. An illustration of the methodology adopted for conducting the survey categorizes depth estimation databases and the loss functions.

![Datasets searching criteria](image)

FIGURE 2. An illustration of database according to the number of citations in each year from 2017 to 2021. The number against each database represents the total citations in each year.

IV. PUBLICLY AVAILABLE DEPTH ESTIMATION DATASETS

This section presents an overview with tabular summaries of the most widely used image depth datasets and classifies them into different use case applications.

Numerous interesting datasets are available for training depth estimation models for both multi-view and monocular images. The datasets general metadata includes details on the number of objects, scenes, and the number of RGB and depth images. The ground truth includes different types of knowledge available in each dataset, including depth, mesh, camera trajectories, video, poses, point cloud, semantic label, trajectory, and dense multi-class labels.

With the growth (evolution) in image depth estimation research, increasing efforts are made in generating larger and more ambitious depth estimation datasets. One growing trend is the increasing number of new publicly available depth estimation datasets becoming available each year over the last ten (10) years. This trend is shown in Fig. 3. A structured taxonomy showing the importance of the depth estimation datasets is given in Fig. 4. The datasets are further divided into different environments (i.e., real/synthetic indoor/outdoor, static indoor/outdoor, and real/generated facial) in Figure 4.

Large and diverse training sets are required for depth estimation.
Since obtaining pixel accurate ground-truth depth at scale in a range of circumstances is challenging, different datasets with specific characteristics and biases have been proposed.

A. THE TYPE AND REPRESENTATIONS OF DATA

There are different types (i.e., alphanumeric, text, image, video, point cloud, mesh, voxel) and representations of data such as (stereo 2D, 2.5D, 3D) that are used to analyse the scenes from different perspectives (e.g., angles).
The most up-to-date depth datasets are divided into many use case applications, such as (camera tracking, scene reconstruction, tracking, semantic, pose, video, streets, people i.e., identity recognition and faces, and medical depth-based applications, indoor and outdoor scenes). A detailed comparative analysis for various data representations is provided in Table 1. Moreover, as some datasets contain data of various types and categories, Table 2 – 11 tabulates a comparative study for the data present in each dataset using the following labels:

- **RGB**: 2-dimensional visible light spectrum images.
- **Depth**: generic term for a map of per-pixel data containing depth-related information. A depth map describes at each pixel the distance to an object (e.g., distance from camera).
- **Video**: sequence of temporally consecutive visual readings.
- **Point cloud**: data composed of a collection of points representing a 3-dimensional shape, where each point has at least an x, y, z coordinate.
- **Mesh**: polygon-based representation of 3-dimensional shapes that directly captures topology and shape surface.
- **Scene**: data recording some environment such as a room.
- **Semantic**: labels mapping some data to a class in some ontology (e.g., human, vehicle, etc.).
- **Object**: data capturing features of objects such as shape or motion. Suitable for tasks such as object classification or tracking.
- **Camera**: data that can be used to track the camera’s geometrical features.
- **Action**: data recording subjects performing certain actions.
- **Trajectory**: data capturing the path of motion or action being performed by some object or entity.
- **Pose**: data specifying human pose information, such as head pose.

### Table 1: Comparison Between Data Representations

| Representation | Data Dimension | Shape Details | Memory Efficiency | Computation Efficiency |
|----------------|----------------|---------------|--------------------|------------------------|
| RGB            | 2D             | ✦ ✦           | ✦                  | ✦                     |
| Depth          | 2.5D           | ✦ ✦           | ✦                  | ✦                     |
| Mesh           | 3D             | ✦             | ✦ ✦ ✦              | ✦                     |
| Voxel          | 3D             | ✦             | ✦ ✦ ✦              | ✦                     |
| Point cloud    | 3D             | ✦             | ✦ ✦ ✦              | ✦                     |
| Octree         | 3D             | ✦             | ✦ ✦ ✦              | ✦                     |
| TSDF           | 3D             | ✦             | ✦ ✦ ✦              | ✦                     |
| Stixel         | 2.5D           | ✦             | ✦ ✦ ✦              | ✦                     |

✦: low; ✦:+: moderate; ✦✦✦: high

### B. Depth datasets for people detection and action recognition

Datasets that capture people doing different tasks like walking and acting as well as human recognition and activity depth datasets can play an important role. By employing depth map people datasets, the goal is to recognize the subject’s identity, gender, or other qualities and activities.

1) **RGB-D PEOPLE**

The RGB-D people dataset [27] contains over 3,000 RGB and depth frames collected from three Kinect sensors mounted vertically in a university hall. The data is comprised of up-right walking and standing humans seen from various angles with various degrees of occlusion. The data is gathered in a middle position (i.e., the lobby of a large canteen) by observing people’s unscripted behaviour during lunch time. The video sequences are captured at 30Hz using a set of three Kinect v1 sensors vertically joined (130° x 50° field of view). This capturing device is around 1.5 meters away from the ground. It ensures that the three images are captured in a synchronized and simultaneous manner while also reducing IR projector crosstalk between the sensors. To reduce sensor biases, certain background samples are taken from another building on the College campus. Occlusions between people is present in most sequences to make the data more realistic. Following the ground truth, all frames are manually annotated with bounding boxes in 2D depth image space and subject visibility position. A total of 1,088 frames, including 1,648 instances of persons, have been labelled to smooth the evaluation of individual detection systems.

2) **TST FALL DETECTION V2**

During the simulation of Activities of Daily Living (ADLs) and falls, the dataset [28] contains depth frames and skeleton joints collected using Microsoft Kinect v2 and acceleration samples provided by an inertial measurement unit (IMU).

The ADLs dataset is simulated for 11 young actors. The actions listed below are included in the ADL category: the actor sits in a chair; the actor walks and grabs an object from the floor; the performer takes a walk back and forth; the actor lies down on the floor. The following actions are included in the category of fall: In the front, the actor falls to the ground...
and lies down; at the back, the actor falls backward and ends up lying; at the side, the actor falls to the side and ends up lying; EUpSit, the actor falls backward and ends up sitting. Each actor performed each action three times, resulting in a total of 264 sequences. The following information is provided for each sequence: Two raw acceleration streams, provided by IMUs constrained to the actor’s waist and right wrist; skeleton joints in depth and skeleton space, captured by Microsoft SDK 2.0; depth frames with a resolution of 512x424, captured by Kinect v2; timing information, timestamps of Kinect frames and acceleration samples, useful for synchronization.

3) WEB STEREO VIDEO
The web stereo video dataset can be used for depth from monocular video sequences containing a large number of non-rigid objects, such as people. To learn non-rigid scene reconstruction cues, [2] includes 553 stereoscopic videos from YouTube. This dataset contains a wide range of scene types as well as several non-rigid features.

4) MANNEQUIN CHALLENGE
In-wild recordings of people in static poses as a handheld camera pan around the environment are available in the mannequin challenge dataset [29]. The dataset is split into three parts for training, validation, and testing. The mannequin challenge is a film collection of people replicating mannequins by freezing in a variety of natural poses as a handheld camera covers the scene. More than 170K frames and associated camera postures were retrieved from around 2,000 YouTube videos in the dataset. SLAM and bundle adjustment techniques were used to calculate the camera poses. The Mannequin Challenge dataset has been used to train the model for predicting dense depth maps from common video with the camera and participants in the scene moving.

5) MHAD
Except for one senior person, the Berkeley Multimodal Human Action Database (MHAD) [30] contains 11 acts done by 7 male and 5 female subjects between the ages of 23 and 30. All the individuals repeated each action five times, resulting in about 660 action sequences and 82 minutes of total recording time. In addition, they recorded a T-pose for each subject which can be used for the skeleton extraction; as well as the background data (i.e., with and without the chair used in some of the activities). Actions with movement in both upper and lower extremities, such as jumping in place, jumping jacks, and throwing; actions with high dynamics in upper extremities, such as waving hands and clapping hands; and actions with high dynamics in lower extremities, such as sitting down and standing up, are included in the specified set of actions. The subjects were given instructions on what action to complete before each recording, but no exact specifics on how the activity should be carried out were supplied (i.e., performance style or speed). As a result, some of the activities have been performed in a variety of styles by the individuals (e.g., punching, throwing). Depth data is collected using two Microsoft Kinect v1 sensors placed in opposite directions to prevent active pattern projection interference.

6) UR FALL DETECTION
The dataset [31] has 70 sequences (30 falls + 40 activities of daily living). Falling events are captured using two Microsoft Kinect v1 cameras and accelerometric data. Only one device (camera) and an accelerometer are used to record ADL actions. PS Move (60Hz) and x-IMU (256Hz) devices were used to collect sensor data.

7) MOBILE-RGBD
On the mobile platform, MobileRGBD is a corpus dedicated to low-level RGB-D dataset [32]. It flipped the traditional corpus recording paradigm on its head. The goal is to make ground truth annotation and record reproducibility easier in the face of speed, trajectory, and environmental changes. To portray static users in the environment, they utilized dummies that do not move between recordings. It is feasible to record the same motion multiple times to validate the impact of detecting algorithms at different speeds. This benchmark corpus is for low-level RGB-D algorithms such as 3D-SLAM, body/skeleton tracking, and face tracking with a mobile robot. Depth data was collected using a Kinect v2 sensor.

| TABLE II |
| --- |
| DEPTH DATASETS FOR PEOPLE DETECTION AND ACTION RECOGNITION. |
| DATASET | RGB | DEPTH | VIDEO | POINT-CLOUD | MESH | SCENE | SEMANTIC | OBJECT | CAMERA | ACTION | TRAJECTORY | POSE |
| RGB-D-P [27] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| TST-F-V2 [28] | ✓ | ✓ | x | x | x | ✓ | ✓ | x | x | x | x | x |
| W-S [2] | ✓ | ✓ | ✓ | x | x | ✓ | ✓ | x | x | x | x | x |
| M-E [29] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x |
| MHAD [30] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x |
| U-F-D [31] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x |
| M-RGBD [32] | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x |

✓: AVAILABLE; x: NOT AVAILABLE; M-E: MANNEQUIN; W-S: WEB STEREO; RGB-D-P: RGB-D PEOPLE; TST-FALL-V2: TST FALL-V2; U-F-D: UR FALL DETECTION; M-RGBD: MOBILE-RGBD.
C. Depth datasets for faces and poses.
Aside from providing a low-cost camera sensor that produces both RGB and depth information, the depth camera sensor also allows a faster human-skeletal tracking. This tracking technique can offer the exact location of human body joints across time, making analyses of complex human behaviours simpler and faster. As a result, deducing human faces from depth images or combining depth and RGB images has received much attention. In recent years, several of these new depth datasets have been developed to help in the verification of human facial activity analysis techniques.

1) BIWI
BIWI dataset [33] with over 15K images of 20 people (6 females and 14 males - 4 people were recorded twice). A depth image, the associated RGB image (both 640x480 pixels), and the annotation are provided for each frame. The range of head poses is approximately +75 degrees yaw and +60 degrees pitch. The ground truth is provided in the form of the head’s 3D location and rotation. Depth data is acquired using a Kinect v1 sensor.

2) EURERCOM KINECT FACE
The multimodal face images of 52 persons (14 females, 38 males) acquired by Kinect v1 are included in the Dataset [34]. The data was collected in two sessions at different times (about half a month). In each session, the dataset provides the facial images of each person in 9 states of different facial expressions, lighting, and occlusion conditions: neutral face, smiling, open mouth, strong illumination, occlusion of eyes by sunglasses, occlusion of mouth by hand, occlusion of side of face by paper, right profile, and left profile. The RGB color image, the depth map (given in two forms of the bitmap depth image and the text file containing the actual depth levels sensed by Kinect), and the 3D image are all produced in three formats. The dataset also includes manual landmarks for six facial positions: left eye, right eye, the tip of the nose, left corner of the mouth, right corner of the mouth, and the chin.

3) PANDORA
The Pandora dataset [35] has 250K full-resolution RGB and depth images, obtained from a Kinect v2 sensor, as well as their annotations. For head centre localization, head pose estimation, and shoulder pose estimation, the Pandora dataset is frequently utilized.

4) FACESCAPE
The FaceScape dataset [36] contains large-scale and high-quality 3D face models, parametric models, and multi-view images. The camera settings, as well as the subjects age and gender, are all included. The information has been made available to the public for non-commercial research purposes. The FaceScape dataset contains 18,760 textured 3D faces, each with 20 distinct expressions, captured from 938 subjects. The pore-level facial geometry is also processed to be topologically uniformed in the 3D models. For rough shapes, these fine 3D facial models can be represented as a 3D morphable model, and for detailed geometry, as displacement maps. Using a deep neural network to learn the expression specific dynamic features, a novel approach is proposed that takes advantage of the large-scale and high-accuracy dataset.

5) 3DMAD
The 3D Mask Attack Database [37] (3DMAD) is a database for spoofing biometric (facial) data. It contains 76500 frames of 17 people captured with Kinect v1 for real-time spoofing attacks. A depth image (640x480 pixels – 1x11 bits), the corresponding RGB image (640x480 pixels – 3x8 bits), and carefully labelled eye positions make up each frame (concerning the RGB image). For each person, data is collected in three separate sessions such that in each session capturing five 300-frame recordings. The recordings are conducted in a controlled environment with a frontal view and neutral expression. The first two sessions are dedicated to real-world samples, in which individuals are recorded with a two-week gap between captures. A single operator captures 3D mask attacks in the third session (attacker).

### TABLE III:
PROPERTIES OF DEPTH DATASETS FOR PEOPLE DETECTION AND ACTION RECOGNITION

| No | Name            | Year | Ground truth type | Size           | Objects per image | Type                  | No images |
|----|-----------------|------|-------------------|----------------|-------------------|-----------------------|-----------|
| 1  | RGB-D-P [27]    | 2011 | Depth             | 640 × 480      | Multiple          | Real, in outdoor      | 3500      |
| 2  | MHAD [30]       | 2013 | depth, 3D position| 3600 × 3600    | Multiple          | Real, body            | 36940     |
| 3  | MOBILE-RGBD [32]| 2014 | Depth             | 1900 × 1080    | Multiple          | Real, indoor          | 36257     |
| 4  | UR FALL DETECTION [31]| 2014 | Depth             | 640 × 480      | Multiple          | Real, indoor          | 7000      |
| 5  | TST FALL-V2 [28]| 2016 | Depth             | 640 × 480      | Multiple          | Real, indoor          | 15000     |
| 6  | WEB STEREO [2]  | 2019 | Depth, 3D models  | 1080 × 1080    | Multiple          | Real, outdoor         | 5000      |
| 7  | MANNEQUIN [29]  | 2019 | Depth             | 640 × 480      | Multiple          | Real, in-outdoor      | 170K      |
1. **KITTl**

KITTl [38] is one of the most often used datasets in mobile robots and self-driving cars. It contains hours of videos of traffic scenarios captured with a range of sensor modalities, including high-resolution RGB and grayscale stereo cameras, as well as a 3D laser scanner (LiDAR). The dataset itself does not contain ground truth for semantic segmentation. However, various researchers have annotated parts of the dataset manually to meet their needs. The authors in [39] created ground truth for 323 images from the road detection challenge, divided into three categories: road, vertical, and sky. The work in [40] annotated 252 (140 for training and 112 for testing) acquisitions, RGB and Velodyne LiDAR scan, from the tracking challenge for ten object categories including building, sky, road, vegetation, sidewalk, car, pedestrian, cyclist, sign/pole, and fence. The authors in [41] labelled 170 images for training and 46 images for testing (from the visual odometry challenge) with 11 classes: building, tree, sky, car, sign, road, pedestrian, fence, pole, sidewalk, and bicyclist.

2. **CITYSCAPES**

The Cityscapes dataset [42] is a large-scale dataset dedicated to the semantic evaluation of urban street scenes. It includes semantic, instance-based, and dense pixel annotations for 30 classes divided into eight groups (i.e., flat surfaces, humans, vehicles, constructions, objects, nature, sky, and void). Around 5,000 finely annotated images and 20,000 coarsely annotated images make up the dataset. The data was collected in 50 places for several months, during daylight hours and under favourable weather circumstances. It was originally shot on video; therefore, the frames were hand-picked to include a large number of dynamic objects, a dynamic scene layout, and a changing background. It also contains 5,000 polygonal annotations, 5,000 volume annotated images for both fine and course annotations, video frames, GPS coordinates, Ego-motion, and outside temperature data from the vehicle sensor and odometry. In terms of diversity, cityscapes are one of the most popular benchmark datasets.

3. **DRIVING STEREO**

DrivingStereo is a large-scale stereo dataset [43] that was created. It is hundreds of times larger than the KITTI stereo dataset, with over 180k images covering a wide range of driving scenarios. A model-guided filtering technique from multi-frame LiDAR points produces high-quality disparity labels. Deep-learning models trained on the DrivingStereo dataset achieve higher generalization accuracy in real-world driving scenes than models trained on other datasets. The dataset contains left and right images along with disparity maps and depth maps. The total number of images 182188 is further divided into 174437 for training and 7751 pairs for testing.

4. **KITTl-DEPTH**

The depth maps from projected LiDAR point clouds were matched against the depth estimation from the stereo cameras in the KITTI-depth dataset [44]. It contains 93K depth maps.
with corresponding raw scene and RGB images captured with LiDAR aligned with the raw KITTI Dataset. On the benchmark server, there are 86k training images, 7k validation images, and 1k test set images. This dataset will enable the training of advanced deep learning models for the problems of depth completion and single image depth prediction.

5) UASOL
The UASOL RGB-D stereo dataset [45] has 160,902 frames captured in 33 separate scenes with between 2k and 10k frames each. The frames represent different pathways, such as sidewalks, trails, and roadways, as seen through the eyes of a pedestrian. The images were extracted from HD2K video files having a resolution of 2280x1282 pixels and a frame rate of 15 frames per second. Each second in the sequences has a GPS geolocation identifier, and the dataset reflects various climatological circumstances. It also involves up to four people photographing the dataset several times during the day.

6) DDAD
DDAD is a new autonomous driving dataset [25] from the Toyota Research Institute (TRI) for long-range (up to 250m) and dense depth estimation in challenging and diverse urban environments. It includes monocular movies as well as accurate ground-truth depth (over a full 360-degree field of view) generated by high-density LiDARs placed on a fleet of self-driving automobiles driving across the United States. Scenes from cities in the United States (San Francisco, Bay Area, Cambridge, Detroit, Ann Arbor) and Japan (Tokyo, Odaiba) appear in DDAD.

7) DENSE
DENSE (Depth Estimation on Synthetic Events) [46] is a novel dataset with pixel accurate ground truth. The camera specifications are set to imitate the MVSEC event camera, which has a sensor size of 346 x 260 pixels and a horizontal field of view of 83 degrees. DENSE is divided into five training sequences, two validation sequences, and one testing sequence. Each sample is a tuple containing one RGB image, the stream of scenes between 2 subsequent images, ground truth depth, and segmentation labels. Each sequence has 1000 samples at 30 frames per second.

8) HEADCAM
This dataset [47] features panoramic video captured while riding a bike around suburban Northern Virginia with a helmet-mounted camera. The videos were used to test an unsupervised learning system for estimating depth and ego motion. The videos are saved as.mkv video files with lossless H.264 compression.

### TABLE VI

PERCEPTION-BASED NAVIGATION DEPTH DATASETS (I.E., STREET SIGNS, ROADS).

| Dataset | RGB | Depth | Video | Point-Cloud | Mesh | Scene | Semantic | Object | Camera | Action | Trajectory | Pose |
|---------|-----|-------|-------|------------|------|-------|----------|--------|--------|--------|------------|------|
| KITTI [38] | ✓   | ✓     | ✓     | ✓          | ×    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| CITYSCAPES [42] | ✓   | ✓     | ✓     | ✓          | ×    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| D-S [43] | ✓   | ✓     | ×     | ×          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| K-D [44] | ✓   | ✓     | ×     | ×          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| UASOL [45] | ✓   | ✓     | ×     | ×          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| DDAD [25] | ✓   | ✓     | ✓     | ✓          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| DENSE [46] | ✓   | ✓     | ✓     | ✓          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |
| H-CAM [47] | ✓   | ✓     | ×     | ×          | ✓    | ✓     | ✓        | ✓      | ✓      | ✓      | ✓          |       |

✓: AVAILABLE; ×: NOT AVAILABLE, D-S: DRIVING STEREO; K-D: KITTI-DEPTH; H-CAM: HEADCAM.

### TABLE VII

PROPERTIES OF PERCEPTION-BASED NAVIGATION DEPTH DATASETS (I.E., STREET SIGNS, ROADS).

| No | Name | Year | Ground truth type | Size | Objects per image | Type       | No images |
|----|------|------|-------------------|------|-------------------|------------|-----------|
| 1  | KITTI [38] | 2012 | Depth             | 1382 × 512 | Multiple | Real, outdoor | 12919    |
| 2  | CITYSCAPES [42] | 2016 | Semantic, instance-wise, depth | 1024 × 2048 | Multiple | Real, outdoor | 25000    |
| 3  | KITTI-DEPTH [44] | 2017 | Depth             | 1382 × 512 | Multiple | Real, outdoor | 94K      |
| 4  | DRIVING STEREO [43] | 2019 | Disparity, depth   | 881 × 400  | Multiple | Real, outdoor | 182188   |
| 5  | HEADCAM [47] | 2019 | Depth             | 512 × 128  | Multiple | Real, outdoor | 27538    |
| 6  | UASOL [45] | 2019 | depth, segmentation, GPS | 2280 × 1282 | Multiple | real, outdoor | 160902   |
| 7  | DENSE [46] | 2020 | Depth             | 346 × 260  | Multiple | Synthetic | 8000     |
| 8  | DDAD [25] | 2020 | 2D, 3D point cloud | 1936 × 1216 | Multiple | Real, outdoor | 71600    |
E. Object and scene recognition depth datasets

Object recognition determines whether the input image contains the pre-defined object, while scene recognition labels all objects in a scene in a dense manner. With the help of object recognition methods, one can distinguish the differences between objects and determine many distortions that might occur such as different occlusions levels, illumination variations, and reflections. Combining RGB and depth information could potentially improve the robustness of the feature methods. Several depth datasets are generated for different tasks in depth object and scene recognition.

1) NYU-D V2

NYU-D V2 [48] is mainly composed of video sequences from a variety of indoor environments captured by the Microsoft Kinect v1 RGB and depth cameras. It consists of 1,449 richly annotated pairs of aligned RGB and depth images from over 450 scenes across three cities. A class and an instance number are assigned to each object (e.g., cup1, cup2, cup3, etc.). There are also 407,024 unlabelled frames in the collection. In comparison to other datasets, this one is relatively small. This dataset was used as a benchmark for indoor depth, segmentation, and classification in the representative study work.

2) SCANNET

ScanNet [49] is an indoor RGB-D dataset that includes both 2D and 3D data at the instance level. Rather than points or objects, it is a collection of labelled voxels. ScanNet v2, the most recent version of ScanNet, has collected 1513 annotated scans with a surface coverage of over 90%. This dataset is divided into 20 classes of annotated 3D voxelized objects for the semantic segmentation challenge.

3) SUN3D

SUN3D includes [50], a large-scale RGB-D video database with 8 annotated sequences. Each frame contains a semantic segmentation of the scene’s features in conjunction with the information on the camera’s position. It is made up of 415 segments captured in 254 distinct locations across 41 different buildings. Furthermore, several locations have been photographed multiple times throughout the day. Depth acquisition was performed using the Asus Xtion Pro Live which utilizes depth from structured light technology.

4) SUN RGB-D

There are 10335 realistic RGB-D images of room scenes in the SUN RGB-D dataset [51]. Each RGB image has a depth and segmentation map that corresponds to it. There are almost 700 different objects with labelled categories. There are 5,285 and 5,050 images in the training and testing sets, respectively. The entire dataset is fully annotated, including 146,617 2D polygons and 58,657 3D bounding boxes with detailed object orientations, as well as a 3D room layout and scene categorization. This dataset allows us to train data-hungry scene-understanding algorithms, evaluate them using direct and relevant 3D metrics, minimize overfitting to a limited testing set, and investigate cross-sensor bias. Four sensors, leveraging three different depth technologies, were used for gathering depth data: Intel RealSense (depth-from-stereo), Kinect v1 and Asus Xtion (structured light), and Kinect v2 (Time-of-Flight).

5) MEGADEPTH

The MegaDepth dataset [52] contains 196 distinct locations reconstructed using COLMAP Structure-from-Motion/Multi-View Stereo (SfM/MVS) for single-view depth prediction. This dataset generates training data from multi-view Internet photo collections, a virtually limitless data source, using sophisticated SfM and MVS algorithms, and presents a large depth dataset named MegaDepth. Data obtained by MVS has its own set of difficulties, such as noise and unreconstructed objects. These issues are addressed by new data cleaning methods, as well as automatically enriching data with ordinal depth relations obtained by semantic segmentation.

6) DIODE

DIODE (Dense Indoor/Outdoor DEpth) [53] is the first standard dataset for monococular depth estimation that includes a variety of indoor and outdoor scenarios captured with the same hardware setup. There are 8,574 indoor and 16,884 outdoor samples in the training set, each with 20 scans. The validation set consists of 325 indoor and 446 outdoor samples obtained from ten separate scans. The indoor training and validation splits have a ground truth density of around 99.54 percent and 99.54 percent, respectively. With 67.19 percent for training and 78.33 percent for validation subsets, the density of the outdoor sets is naturally lower. The datasets ranges are 50m and 300m indoors and outdoors, respectively. Depth data is acquired using the FARO LiDAR.

7) MIDDLEBURY

The Middlebury Stereo dataset [54] contains pixel-accurate ground-truth disparity data and high-resolution stereo sequences with complicated geometry. The ground-truth disparities are obtained using a unique technique that uses structured illumination and does not require the light projectors for calibration. The Middlebury dataset, which contains 38 realistic indoor scenes taken through a structured light scanner, was one of the first datasets for stereo matching. A modified version of the Middlebury dataset with 33 new indoor scenes presented to provide a more accurate annotation at a resolution of 6 Megapixels. They are, however, generally small in size due to the difficulty and expensive cost of creating such exact and dense stereo datasets, which also leads to the problem of low variability. In an indoor setting with controlled lighting, the scenes are limited.

8) EDEN

EDEN (Enclosed garDEN) is a synthetic multimodal dataset for nature-oriented applications [55]. More than 300,000 images were captured from more than 100 garden models in the dataset. Semantic segmentation, depth, surface normals,
intrinsic colours, and optical flow are among the low/high level vision modalities labelled on each image.

9) INRIA DLFD
The INRIA Dense Light Field Dataset (DLFD) [55] is a light field dataset for testing depth estimation methods. There are 39 scenes in DLFD with a disparity range of [-4,4] pixels. The light fields have a 512 × 512 spatial resolution and a 9 × 9 angular resolution.

10) SUNCG
The SUNCG dataset [56] contains 45,622 scenes with realistic room and furniture layouts that were generated manually using the Planner5D platform. Planner5D is a web-based interior design tool that lets users construct multi-floor room layouts, add furniture from a library, and arrange it in the rooms. After deleting duplicated and empty scenes, a simple Mechanical Turk cleaning operation was used to improve the data quality. During the work, the authors display a set of top view renderings of each level and ask the participants to vote on whether or not this is a valid apartment floor. They take three votes for each floor, and a floor is considered valid if it receives at least two positive votes. They have 49,884 valid floors, 404,058 rooms, and 5,697,217 object instances from 2,644 unique object meshes containing 84 categories in the end. They also manually assigned category labels to all the library items.

11) STANFORD 2D-3D
The Stanford 2D-3D dataset [49] collects mutually registered modalities from 2D, 2.5D, and 3D domains, as well as instance-level semantic and geometric annotations, across six indoor areas. It includes more than 70,000 RGB images, as well as depths, surface normals, semantic annotations, global XYZ images, and camera information. Depth data was collected using the Matterport camera, which combines 3 structured-light sensors at different pitches to capture 18 RGB and depth images during a 360° rotation at each scan location.

12) MATTERPORT3D
The Matterport3D dataset [57] is a big RGB-D dataset that can be used to analyze scenes in indoor areas. It is made up of 194,400 RGB-D images and features 10,800 panoramic views inside 90 real building-scale sceneries. Surface construction, camera postures, and semantic segmentation are all annotated in each scene, of a residential building with many rooms and floor levels. The Matterport camera is also used for this dataset.

13) TASKONOMY
Taskonomy [58] offers a vast and high-quality dataset of various indoor environments. This dataset contains comprehensive pixel-level geometry information via aligned meshes, as well as semantic information, derived from ImageNet, MS COCO, and MIT Places, camera positions, complete camera intrinsic parameters, and high-quality images, making it three times the size of ImageNet. This is accomplished by searching a latent space for (first and higher order) transfer learning dependencies across a dictionary of twenty-six 2D, 2.5D, 3D, and semantic tasks.

14) ETH3D
ETH3D is a MVS benchmark/3D reconstruction benchmark that covers a wide range of indoor and outdoor environments [4]. A high-precision laser scanner was used to generate ground truth geometry. Images were captured using a DSLR camera and a synchronized multi-camera system with variable field-of-view. Instead of carefully constructing scenes in a controlled laboratory environment as in Middlebury, ETH3D provides the full range of challenges of real-world photogrammetric measurements. However, it still suffers from a lack of data samples and variability.

15) 2D-3D MATCH
The 2D-3D Match dataset [59] is a novel 2D-3D correspondence dataset that takes advantage of the availability of various 3D datasets from RGB-D scans. The data from SceneNet and 3DMatch are specifically utilised. There are 110 RGB-D scans in the training dataset, with 56 images from SceneNet and 54 scenes from 3DMatch. The following is how the 2D-3D correspondence data is generated. A set of 3D patches from various scanning viewpoints is extracted from a 3D point randomly sampled from a 3D point cloud. Each 3D patch’s 3D position is re-projected into all RGB-D frames for which the point lies in the camera frustum, taking occlusion into consideration, to find a 2D-3D correlation. Around the re-projected point, the matching local 2D patches are extracted. Around 1.4 million 2D-3D correspondences are collected in total.

16) 3D60°
360° [60] repurposed newly released large scale 3D datasets, rendering them to 360°, and creating high-quality 360° datasets with ground truth depth annotations. 3D60 is a collection of datasets created as part of multiple 360° vision research projects (Matterport-3D, Stanford 2D-3D, SunCG). It consists of multi-modal stereo representations of scenarios generated from large-scale 3D datasets, both realistic and synthetic.

17) MINNAV
MinNav is a synthetic dataset based on the sandbox game Minecraft [61]. To generate rendered image sequences with time-aligned depth maps, surface normal maps, and camera poses, the dataset employs multiple plug-in applications. Because of the big gaming community, there is an extremely large number of 3D open-world environments where players can identify acceptable shooting locations and create data sets, as well as create scenes in-game. Sildur renders 300 monocular color images for each camera trajectory, which are stored as 8-bit PNG files with lossless compression. The fps is being adjusted from 10 to 120 and render at 800x600 with fov=70 and fps=10.

18) MAKE3D
The Make3D dataset [62] is a monocular depth estimation dataset with 400 single training RGB and depth map pairs and 134 test samples. While the RGB images have a high resolution, the depth maps have a low resolution of 305×55 generated from a custom 3D laser scanner.

19) TUM RGB-D

TUM RGB-D [63] is an RGB-D indoor dataset that contains colour and depth images from a Microsoft Kinect v1 sensor along with the sensors ground-truth trajectory. The data was captured at a full-frame rate (i.e., 30 Hz) and with a sensor resolution of 1 megapixel (i.e., 640×480). A high-accuracy motion-capture system with eight high-speed tracking cameras provided the ground-truth trajectory (i.e., 100 Hz).

### TABLE VIII

**OBJECT AND SCENE RECOGNITION DEPTH DATASETS.**

| DATASET     | RGB | DEPTH | VIDEO | POINT-CLOUD | MESH | SCENE | SEMANTIC | OBJECT | CAMERA | ACTION | TRAJECTORY | POSE |
|-------------|-----|-------|-------|-------------|------|-------|----------|--------|--------|--------|------------|------|
| NYU-v2[48]  | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| SCANNET [49]| ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ×      | ×      | ✓      | ✓          | ✓    |
| SUN3D [50]  | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ×      | ×      | ✓      | ✓          | ✓    |
| SUN RGB-D [51]| ✓ | ✓ | ✓ | × | × | ✓ | ✓ | × | × | ✓ | ✓ | ✓ |
| M-D [52]    | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ×      | ×      | ✓      | ✓          | ✓    |
| DIODE [53]  | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ×      | ✓          | ✓    |
| MB [54]     | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ×      | ×      | ✓      | ✓          | ✓    |
| EDEN [55]   | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ×      | ✓          | ✓    |
| I-D [55]    | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ✓      | ×      | ×      | ✓          | ✓    |
| SUNCG [56]  | ✓   | ✓     | ✓     | ×           | ×    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| S-2-3D [49]| ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| M-3D [57]   | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| TASKOMONY [58]| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ |
| ETH3D [4]   | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| 2-3D [59]   | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| 360°[60]    | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| MINNAV [61] | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| TUM-D [63]  | ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |
| MAKE3D [62]| ✓   | ✓     | ✓     | ✓           | ✓    | ✓     | ✓        | ✓      | ×      | ✓      | ✓          | ✓    |

✓: AVAILABLE; ×: NOT AVAILABLE; S-2-3D: STANDFORD 2D-3D; MB: MIDDLEBURY; M-3D: METTERPORT3D; M-D: MEGADEPTH, 2-3D:2D-3D MATCH, I-D: INRIA DLF D

### TABLE IX

**PROPERTIES OF OBJECT AND SCENE RECOGNITION DEPTH DATASETS.**

| No | Name              | Year | Ground truth type | Size          | Objects per image | Type          | No images |
|----|-------------------|------|-------------------|---------------|-------------------|---------------|-----------|
| 1  | MAKE3D [62]       | 2008 | depth             | 640 × 480     | Single            | Real, outdoor | 400       |
| 2  | TUM RGB-D [63]    | 2012 | depth             | 640 × 480     | Single            | Real, indoor  | 1510      |
| 3  | NYU-V2[48]        | 2012 | Dense depth       | 640 × 480     | Multiple          | Real, indoor  | 1449      |
| 4  | SUN3D [50]        | 2013 | depth, semantic, 3D Gt | 640 × 480 | Multiple          | Static, indoor | depth.3D  |
| 5  | MIDDLEBURY [54]   | 2014 | depth             | 1080 × 1080   | Multiple          | Static, indoor | 8640      |
| 6  | SUN RGB-D [51]    | 2015 | depth, semantic, 2D and 3D | 640 × 480 | Multiple          | Static, indoor | 10353     |
| 7  | SCANNET [49]      | 2017 | labelled voxels, depth, 2.5depth | 1920 × 1080 | Multiple          | Synthetic, indoor | 2.5M       |
| 8  | SUNC G [56]       | 2017 | 2D, 3D, volumetric, Depth | 256 × 160 | Multiple          | Synthetic, Indoor | 45622     |
| 9  | STANDFORD 2D-3D [49]| 2017 | 2.5 depth, meshes, point cloud | 1080 × 1080 | Multiple          | Real, indoor  | 70496     |
| 10 | METTERPORT3D [57] | 2017 | depth, 2D, 3D semantic | 1280 × 1024 | Multiple          | Synthetic, indoor | 194400    |
F. Depth datasets for medical applications

In the last decade, medical recognition utilizing depth maps has seen significant research. As a result, depth maps-based medical methods are being employed for various applications, including monitoring of radiation in image-guided interventions to decrease surgical stuff exposure to X-rays, endoscopic surgeries for real time safety monitoring, and navigation analysis to support ultrasound procedures. Various datasets have been generated to address different medical task-based applications.

1) EndoSLAM

The endoscopic SLAM dataset [64] (EndoSLAM) is a dataset for endoscopic video depth estimation. This includes 3D point cloud data for six porcine organs, capsule and standard endoscopy recordings, synthetically produced data, and clinically used conventional endoscopy recordings of the phantom colon with computed tomography (CT) scan ground truth.

2) MVOR

The Multi-View Operating Room (MVOR) dataset [65] consists of 732 multi view frames captured by three RGB-D cameras (Asus Xtion Pro). Every frame consists of three RGB and depth images. The data was sampled from four days of recording in room at the hospital during vertebroplasty and lung biopsy. There are in total 2,926 2D key point annotations, 4,699 bounding boxes and 1,061 3D key point annotations.

3) Cholec80

The Cholec80 dataset [66] consists of 80 videos for cholecystectomy surgeries performed by different surgeons.

The videos were shot at a frame rate of 25 frames per second. The timing (at 25 frames per second) and tool presence annotations are included in the dataset (at 1 fps). The dataset is divided into two equal-sized subgroups (i.e., 40 videos each). There are around 86K annotated images in the first subset. Ten videos from this selection have also been thoroughly annotated with tool bounding boxes. The evaluation subgroup (the second subset) is utilized to put the algorithms for tool presence detection and phase recognition to the test.

4) xawAR16

The xawAR16 dataset [67] is multi-view RGB-D camera dataset that was created in an operating room (IHU Strasbourg) to test the tracking and relocalization of a hand-held moving camera. To create such a dataset, three RGB-D cameras (Asus Xtion Pro Live) were employed. Two of them are fixed to the ceiling in such a way that they may capture views from both sides of the operating table. A third is attached to a display that is moved around the room by a user. A moving camera is fitted with a reflecting passive marker, and its ground-truth pose is determined using a real-time optical 3D measuring system. The dataset consists of 16 time-synchronized color and depth images in full sensor resolution (640x480) captured at 25 frames per second, as well as ground-truth positions of the moving camera measured at 30 frames per second by the tracking device. Each sequence includes occlusions, motion in the scene, and sudden perspective shifts, as well as varied scene layouts and camera movements.

| DATASET       | RGB | DEPTH | VIDEO | POINT-CLOUD | MESH | SCENE | SEMANTIC | OBJECT | CAMERA | ACTION | TRAJECTORY | POSE |
|---------------|-----|-------|-------|-------------|------|-------|----------|--------|--------|--------|------------|------|
| EndoSLAM [64] | ✓   | ✓     | ✓     | ✓           | X    | ✓     | X        | ✓      | ✓      | ✓      | ✓          | ✓    |
| MVOR [65]     | ✓   | ✓     | ✓     | ✓           | X    | ✓     | X        | ✓      | ✓      | ✓      | ✓          | ✓    |
| Cholec80 [66] | ✓   | ✓     | ✓     | ✓           | X    | ✓     | X        | X      | ✓      | ✓      | ✓          | X    |
| xawAR16 [67]  | ✓   | ✓     | ✓     | ✓           | X    | ✓     | X        | X      | ✓      | ✓      | ✓          | X    |

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
G. Explanation and datasets comparison
This section demonstrate brief comparison of depth datasets from several aspects. For an easy access, all the datasets are ordered by year; table 6 shows some features including the name of the datasets, the year of creation, ground truth type, size, objects per image in the dataset, type, and number of images. In terms of popularity of the datasets, the authors ranked the datasets based on the number of citations. The datasets that are available freely and with longer history always have more citations than the newer ones. Particularly Kitt, Cityscapes, Nyu-v2, Sun-RGB-D, Make3D, SceneNet, SunCG all have high number of citations compared to the rest of the datasets. However, it does not necessarily mean that the old datasets are better than the new ones. In terms of the baseline evaluation datasets for depth estimation, Kitti, Cityscapes, Nyu-v2 are the commonly used benchmarks. The depth datasets are divided into different categories of intended applications and studied properties. However, each dataset may not be limited to one specific application only (e.g. Kitti can be used for both depth and 3D reconstruction, Nyu-v2 can be used for both depth and segmentation). The data modalities include RGB, depth, indoor, outdoor, real, synthetic, semantic, labeled voxels, 3D, volumetric, meshes, point cloud, 3D landmarks, surface normals, camera poses, and segmentation. This is helpful for researchers to quickly identify the datasets of interest especially when they are working on multi-modal fusion. A link to each dataset is also provided, which can help research involved in similar studies. It is important to keep in mind that some datasets are updated while others’ websites may change.

H. Mixing Datasets for Training on Diverse Data
To the author's knowledge, the systematic combination of many data sources has only been briefly studied. [68] described a model for estimating two-view structure and motion, which they trained on a combination of smaller datasets with static scenes; although, they did not explain the impact of the method used. [69] proposed a method of naïvely mixing datasets for monocular depth estimation with known camera parameters. Combining different datasets can be challenge as the ground truth data is in different forms (i.e., absolute form: laser based or stereo camera with unknown camera parameters, depth from unknown scale, disparity maps) in every dataset (see table 3). A methodology that can be compatible with all ground truth representations for training deep networks is required. Furthermore, an appropriate loss function can be designed, which must be flexible and compatible with different kind of ground data sources.

Three key issues are identified by [10] and studied in detail.

• Direct vs. inverse depth representations are inherently different representations of depth.
• Scale ambiguity: depth with unknown scale (or camera parameters, camera calibration) in some data sources.
• Uncertainty about shift: some datasets only include disparity maps up to a certain known scale.

Although a stochastic optimization computation, loss function and prediction space allow for the mixing of different data sources, while it is not instantly obvious in what percentages different datasets will be merged through training.

When it comes to mixing datasets, there are two crucial approaches to consider.

1. In each minibatch, the first technique is to combine different data sources into equal parts which sample F/K training data from each dataset for a minibatch of size F, where K specifies the number of different datasets. This technique ensures that all datasets, regardless of the size, are characterized equally in the effective training set for training deep networks.

2. The second approach takes a more principled style, adapting a recent Pareto-optimal multi-task learning method [70]. They examine every dataset as a different task and try to find an approximated Pareto optimum across all datasets (i.e., a technique in which the loss on each training set cannot be reduced without raising it on at least one of the others). To minimize the multi-objective optimization criteria, it utilizes the algorithm provided in [70] that can be used for mixing different kind of ground truth data into an effective way for various tasks in computer vision-based applications.

\[
\min_f (L_1(f), ..., L_t(f))^t
\]

Where parameters of the model \( f \) are shared across different datasets.

V. COMMON CHARACTERISTICS OF WELL-KNOWN DATASETS
It was observed that, of the datasets mentioned above, the depth estimation datasets with the highest potentials displayed five common qualities:

• Longevity -This study finds that the datasets that were available for a longer period of time gained more attention and popularity. The KITTI is the most discussed dataset and
has been accessible since its launch in 2012. It is the most frequently cited benchmark dataset despite several constraints, such as small scale. The KITTI dataset has become a standard benchmark for comparing new results and methods for depth estimation and 3D reconstruction tasks.

- Scale – The number of samples and subjects in a dataset plays a critical role in its popularity. A dataset must have enough sample data features for successful statistical research. Datasets with many samples (and thus a higher statistical relevance) provide objective standards. In conjunction with the dataset size, some other features such as the methodology of its representation are also important.

- Timing – It is observed that the most popular datasets provided novel features and facilitated research that was not possible with previously available public datasets. The KITTI dataset, which was the first publicly available depth outdoor dataset, the NYU-V2 dataset, which was the first dataset to add indoor imaging, and the Cityscapes dataset, which was the first to feature high-resolution images, are all good examples.

- Data quality - The data quality plays a critical role in providing the information about its use in the given situation (e.g., data analysis). It is worth noting that the datasets with details for information collection usually get more attention than the rest of the datasets (e.g., NYU-D V2, FaceScape, Cityscapes).

- The Right Data Transformation - Once generated, the datasets are modified for meeting particular performance objectives while using the machine learning algorithms. Domain knowledge and algorithm features/functions can help determine the best type of transformation to increase the training performance. Datasets that include tools for cleaning, transforming, and preparing data for training are popular than research-oriented datasets.

### VI. State-Of-The-Art Depth Estimation Methods On Three Most Popular Datasets

The performance of the top five SoA algorithms on popular depth estimation benchmarks is tabulated in this section. It’s worth noting that, while most deep networks report their results using standard datasets and metrics, some don’t, making it impossible to compare SoA methods across the board. Furthermore, only a small percentage of papers provide reliable additional information, such as execution time and memory footprint, which is critical for industrial depth estimation model applications (such as drones, self-driving cars, robotics, and so on) that must run on embedded consumer devices with limited processing power and storage and thus require efficient, lightweight models. The performance of the top five SoA deep learning-based depth estimation models on three of the most popular datasets is summarized in Tables 12-14.

| Method               | Dataset         | RMS      | Year |
|----------------------|-----------------|----------|------|
| 3d-ken-burns [71]    | NYU-V2          | 0.305    | 2019 |
| AdaBins [72]         | NYU-V2          | 0.364    | 2020 |
| TransDepth [73]      | NYU-V2          | 0.365    | 2021 |
| BTS [74]             | NYU-V2          | 0.407    | 2019 |
| Optimized, freeform  [75] | NYU-V2          | 0.432    | 2019 |

### VII. An Overview Of Loss Functions For Depth Estimation

Deep learning-based methods usually optimize a regression model on the reference depth map. For depth regression tasks, defining an appropriate loss function is the main challenge faced by the SoA methods. Optimisation algorithms are used by neural networks (i.e., stochastic gradient descent to minimize the errors in the algorithm). The loss function, which measures how well or poorly the model performs, is used to calculate this error. There are several noteworthy loss functions that have been employed in depth estimation problems where deep neural networks are used to forecast depth maps from a single or multiple images.

#### A. Least Square Loss

To supervise the training process of the models, the differences between the real depth $y$ and predicted $\hat{y}$ maps are used. For the depth values, the $L_2$ loss function [73] can be represented as $(L_2)$ and is defined as:

$$ L_2(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 $$ \hspace{1cm} (1)

As a result, depth estimation architectures predict the ground truth to learn the depth information of the scenes.

#### B. Scale-Invariant Loss

During the training stage, depth estimation approaches use the ground truth of depth $y$ and the corresponding model predicts the log depth. The training Scale-invariant loss
function [73] \( L_{SL} \) can be represented by \( L_{aq} \) for the depth values and is defined as:

\[
L_{aq}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \left( \log(y_i) - \log(\hat{y}_i) \right)^2
\]

\[
-\frac{\lambda}{N} \left( \sum_{i=1}^{N} \log(y_i) - \log(\hat{y}_i) \right)^2
\]  

(2)

\( \lambda \) refers to the balance factor and is set to 0.5.

C. BERHU LOSS

To account for data that contains outliers or heavy-tailed errors, the Ordinary Least Square (OLS) estimator is deemed ineffective in this scenario. In the case of Gaussian noise, however, Berhu loss is designed to keep good qualities. Furthermore, the adaptive Berhu penalty encourages a grouping effect, which develops one group with the highest coefficients. Berhu loss function [74] \( L_{berhu} \) can be represented by \( L_{Berhu} \) for the depth values and is defined as:

\[
L_{Berhu}(y, \hat{y}) = \begin{cases} 
(y_i - \hat{y}_i) & \text{if } (y_i - \hat{y}_i) \leq c, \\
(y_i - \hat{y}_i)^2 + c^2 & \text{if } (y_i - \hat{y}_i) > c
\end{cases}
\]  

(3)

D. HUBER LOSS

It is known that Mean Square Error (MSE) is better for learning outliers in a dataset but Mean Absolute Error (MAE) is better for ignoring them. However, data that appears to be outliers should not be considered in some circumstances, and those points should not be given great attention. For this reason, Huber loss function [74] \( L_{huber} \) can be represented by \( L_{Huber} \) for the depth values and is defined as:

\[
L_{Huber}(y, \hat{y}) = \begin{cases} 
(y_i - \hat{y}_i) & \text{if } (y_i - \hat{y}_i) \geq c, \\
(y_i - \hat{y}_i)^2 + c^2 & \text{if } (y_i - \hat{y}_i) < c
\end{cases}
\]  

(4)

E. SILOG LOSS

Correctly scaling the range of the loss function can increase convergence and training outputs, while increasing the \( \lambda \) forces more focus on minimizing the error variance, resulting in Silog loss function [74]. \( L_{silog} \) can be represented by \( L_{silog} \) for the depth values, \( \lambda = 0.5 \) and \( N \) represent ground truth values (i.e., the number of pixels).

By rewriting equation. 2:

\[
L_{silog}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \left( \log(y_i) - \log(\hat{y}_i) \right)^2
\]

\[
-\frac{1}{N} \sum_{i=1}^{N} \left( y_i - \hat{y}_i \right)^2 + (1 - \lambda) \frac{1}{N} \sum_{i=1}^{N} \left( y_i - \hat{y}_i \right)^2
\]

In log space, variance and weighted squared mean errors is combined define the Silog loss:

\[
L_{silog}(y, \hat{y}) = \alpha \sqrt{L_{silog}(y, \hat{y})}
\]  

(5)

F. COMMON DEPTH LOSS

Let \( y \) be a ground-truth depth map and \( \hat{y} \) be its estimated depth. The common depth loss [84] \( L_c \) is given by the entry-wise \( L_1 \)-norm for a matrix

\[
L_c(y, \hat{y}) = \frac{1}{HW} \left( y_i - \hat{y}_i \right)
\]  

(6)

Where \( W \) and \( H \) are the width and height of the depth maps.

G. GLOBAL MEAN REMOVED LOSS

The global mean removed loss [84] is defined as

\[
L_{GMR}(y, \hat{y}) = \frac{1}{HW} \left( (y_i - \bar{y}) - (\hat{y}_i - \bar{\hat{y}}) \right)
\]  

(7)

Where \( W \) and \( H \) are the width and height of the depth maps, \( \bar{y}_i \) and \( \bar{\hat{y}}_i \) are the average depths in \( y \) and \( \hat{y} \), respectively. This loss is based on the observation that, while estimating the global depth scale (i.e., average depth) from an image is unclear, predicting the relative depth of each pixel in relation to the average depth is more reliable. In some situations, such as age estimation, relative estimation is easier than absolute estimation.

H. LOCAL MEAN REMOVED LOSS

A local mean removed loss [84] \( L_{MRL} \), which penalizes the relative depth errors with respect to local \( n \times n \) square regions and defined as follows:

\[
L_{MRL}(y, \hat{y}) = \frac{1}{HW} \left( (y_i - \bar{y}) \oplus J_m - (\hat{y}_i - \bar{\hat{y}}) \oplus J_m \right)
\]  

(8)

Where \( \oplus \) denotes the convolution, and \( J_m \) is the \( n \times n \) matrix composed of all ones.

I. SSIM LOSS

The perceptual difference between two comparable images is measured using SSIM. It can’t tell which of the two is superior because it doesn’t know which is the "original" and
which has undergone further processing like data compression. The loss function for the structural similarity index measure (SSIM) is represented by \((L_{SSIM})\) and can be defined as:

\[
L_{SSIM}(y, \bar{y}) = \left(1 - \frac{L_{SSIM}(y, \bar{y})}{\text{MaxDepth}}\right)
\]

\[(9)\]

**J. PHOTOMETRIC LOSS**

A SSIM term is combined with the \(L_i\) reprojection loss due to its better performance in complex illumination scenarios. Thus, the \((L_p)\) photometric loss \([85]\) of the N scale is modified as

\[
L_p(y, \bar{y}) = \sum_{i}^{N}(1-\lambda)(y_i - \bar{y}_i) + \lambda \frac{1 - L_{SSIM}(y, \bar{y})}{2}
\]

\[(10)\]

**K. PRE-Pixel SMOOTHNESS LOSS**

A per-pixel smoothness loss is introduced to combine with the \(L_{SL}\) reprojection loss to encourage the inverse depth prediction to be locally smooth, as depth discontinuities often occur at image gradients. Thus, the \((L_{SL})\) loss is defined as

\[
L_{SL}(y, \bar{y}) = \sum_{i}^{N} \partial_x dte^{-\partial_x^2(y, \bar{y})} + \partial_y dte^{-\partial_y^2(y, \bar{y})}
\]

\[(11)\]

**L. RECONSTRUCTION LOSS**

The network calculates disparity during training, and the bilinear sample is used to generate the input image, which is then used to reconstruct another image using the disparity map. The bilinear sampler is fully differentiable at the local level and smoothly integrates into a fully convolutional architecture. A \(L_{Huber}\) and SSIM is combined as a photometric image reconstruction loss, which computes the inconsistency between the input image and the reconstructed image, it is defined as follows

\[
L_R(y, \bar{y}) = \frac{1}{N} \sum_{i}^{N} \frac{1 - L_{SSIM}(y_i, \bar{y}_i)}{2} + (1 - \alpha)L_{Huber}((y, \bar{y}))
\]

\[(12)\]

**M. PRIOR RECONSTRUCTION LOSS**

It is consequently shown that constraining a cost function involving a polarimetry-specific geometry is valid. Furthermore, because it is dependent on both the input and output of the processing pipeline, this minimization strategy can be used to optimize a deep learning model. This method is consistent in unusual circumstances, implying a limited camera calibration or a specific azimuth to angle of polarization thought processes. As a result, a new method provides an alternative but comparable strategy that allows for standard calibration and the release of constraints via a generalized loss term defined as follows

\[
L_{PR}(y, \bar{y}) = \mu \min L_R + \psi^2 dte^{-\psi^2(y, \bar{y})} + \partial_y^2 dte^{-\partial_y^2(y, \bar{y})}
\]

\[(13)\]

**N-1. SCALE INVARIANT LOSS**

The scale-invariant loss \([32]\) for a single sample is defined as

\[
L_{SI}(y, \bar{y}) = \frac{1}{N} \sum_{i}^{N} \rho^2(y_i, \bar{y}_i) - \frac{\lambda}{n^2} \left(\sum_{i}^{N} \rho(y_i, \bar{y}_i)\right)
\]

\[(14-1)\]

where \(\rho\) function defines the scale invariant loss and \(\lambda \in [0,1]\).

**N. SCALE SHIFT INVARIANT LOSS**

The scale-shift-invariant loss for a single sample is defined as

\[
L_{SSI}(y, \bar{y}) = \frac{1}{2N} \sum_{i}^{N} \rho(y_i, \bar{y}_i)
\]

\[(14)\]

where \(\rho\) function defines the scale invariant loss.

**O. POINT-WISE LOSS**

Point-wise loss function \((L_{depth})\) can be represented by \((L_1)\) for the depth values and is defined as:

\[
L_{depth}(y, \bar{y}) = \frac{1}{n} \sum \rho_i(y_i - \bar{y}_i)
\]

\[(15)\]

**P. GRADIENT LOSS**

To capture the local structural consistency, a gradient loss function \((L_{grad})\) is proposed and can be represented by \((L_{grad})\), which penalize the gradient of depth around the edges of the image and can be defined as

\[
L_{grad}(y, \bar{y}) = \frac{1}{n} \sum_{i}^{n} \rho_i(y_i(e_i) + \bar{y}_i(e_i))
\]

\[(16)\]

where \(y_i(e_i)\) and \(\bar{y}_i(e_i)\) represent the spatial derivatives of the difference between the ground truth and predicted depth for the \(p^{th}\) pixels \(e_i\) which stands \(\|y_i - \bar{y}_i\|\) for the x, y-axis.
Q. SURFACE NORMAL LOSS

The surface normal loss function \( L_{SN} \) can be utilized to avoid minor errors and predicts the normal and estimated depth maps. The ground-truth surface normals and predicted depth are represented by
\[
n_i^y = (\Psi[-\nabla_x(y_i), -\nabla_y(y_i), 1]^T)
\]
and
\[
n \hat{y}_i = (\Psi[-\nabla_x(\hat{y}_i), -\nabla_y(\hat{y}_i), 1]^T)
\]
The loss is calculated as the difference between the two surfaces normals, which may be expressed mathematically as follows
\[
L_{SN} = \frac{1}{n} \sum_i^n (1 - \frac{\langle n_i^y, n \hat{y}_i \rangle}{\|n_i^y\| \|n \hat{y}_i\|})
\]
where \( \langle n_i^y, n \hat{y}_i \rangle \) denotes the inner product of the vectors.

R. PERCEPTUAL LOSS

The ability of the MSE function to capture perceptually relevant differences (such as high texture details). It is very limited in the use cases because they are defined based on differences in image pixels, minimizing the pixel averages. Therefore, a perceptual loss function is introduced to make the two more perceptible similarities by comparing feature maps between original view and reconstructed view. Denote by \( \alpha \) the feature map obtained after the j-th convolution (after activation) of the i-th convolutional layer in the VGG-16 network and the perceptual loss is defined as the Euclidean distance between the feature maps of the original view \( y \) and the reconstructed view \( \hat{y} \)
\[
L_{PRL}(y, \hat{y}) = \frac{1}{HW} \sum_{i}^N (\alpha(y_i) - \alpha(\hat{y}_i))^2
\]
The size of the generated feature map for a specific layer in the VGG network is described by \( H \) and \( W \). Perceptual loss, rather than pixel-by-pixel loss, is more reflective of semantic similarity between images during training. By adding perceptual loss training, the depth map generated by the model has more precise details and edge information.

S. STRUCTURE GUIDED RANKING LOSS

Structure-Guided Ranking Loss is a pair-wise ranking loss that is very broad, allowing it to be applied to a wide range of depth and pseudo-depth data. The sampling method for certain point pairs, on the other hand, might have a significant impact on the reconstruction quality. Rather than utilizing random sampling, the proposed segment-guided sampling technique and purpose is to direct the networks attention to the regions that matter most, i.e., the scene’s salient depth structures, and can be characterized as
\[
L_{SGL}(y, \hat{y}) = \frac{1}{N} \sum_{i}^N (\alpha(y_i) - \alpha(\hat{y}_i)) + L_{GRAD}(y, \hat{y})
\]

T. CHAMFER LOSS

The chamfer distance between two points can be defined as
\[
D(X_1, X_2) = \sum_{x \in X_1} \min_{y \in X_2} \|x - y\|^2 + \sum_{y \in X_2} \min_{x \in X_1} \|x - y\|^2
\]
for a distance \( d \) between subsets in \( R^2 \). Then the Chamfer loss function takes the form
\[
L_{CL}(y, \hat{y}) = \sum_{i}^N d(y_i, \hat{y}_i)
\]
where \( i \) indexes training samples.

U. BIN CENTER DENSITY LOSS

Bin centre density loss function can be used to follow the distribution of the depth pixels in the ground truth, and it can be defined as the set of bin centres \( c(b) \) and a set of the ground truth pixels in the image \( X \) along with bi-directional Chamfer loss as a regularizes
\[
L_{BCDL} = \sum_{x \in X} \min_{y \in c(b)} \|x - y\|^2 + \sum_{y \in c(b)} \min_{x \in X} \|x - y\|^2
\]

V. GRADIENT MATCHING LOSS

To encourage the network to output a depth map with sharp edges, gradient matching loss is used and defined as
\[
L_{GML}(y, \hat{y}) = \frac{1}{K} \sum_{k=1}^K \sum_{i=1}^K \left( \| \nabla_x E \|_k + \| \nabla_y E \|_k \right)
\]
Where \( \nabla_x \) and \( \nabla_y \) are the gradient of the prediction.

W. PAIRWISE DISTILLATION LOSS

The pairwise distillation loss is obtained in two steps. First, affinity maps for the feature maps are generated. Then the MSE between the affinity maps of the obtained features is then computed.
\[
L_{PDL}(y, \hat{y}) = \frac{1}{x} \sum_{y \in Y} \sum_{i}^N (p_{ij}' - p_{ij}^u)
\]
Where \( p_{ij}' \) and \( p_{ij}^u \) are the affinity maps.

VIII. DISCUSSION

Over the previous two decades, available depth estimation datasets have improved, yet there are still problems to be solved. The most significant limitation is their availability, which implies that many of the datasets are only available for a limited duration. It's also worth noting that in some
circumstances, when the authors prefer to give the dataset based on the asking institutions, limited access is noticed (institutions with a lower profile might typically have more problems obtaining a dataset). This negatively impacts individual researchers’ ability to replicate the analysis, as well as future researchers’ capabilities to publish findings derived from such datasets. The impact of aging has been studied using public datasets collected in the previous few years. Long and complex depth estimation is limited by the difficulties of following up on a large group of people over a long period of time.

The new data privacy standards, which secure personal rights, have created a relatively new challenge. In Europe, for example, the General Data Protection Regulation (GDPR) includes a right to erasure (often known as the right to be forgotten), which gives subjects the option to withdraw their consent to the use of their data and have subject-related material removed from datasets (if possible). Because of the nature of biometric data, the subject can be uniquely identified. As a result, potential changes in datasets could compromise the determination and uniformity of reported data over time. Similar legislations are being discussed globally as a result of recent difficulties relating to the lack of realistic data. Imperfections in the mentioned collection setup and technique are also significant limitations of the current datasets. Some of the dataset generation criteria are not available, but they may be useful so that others can greatly expand the datasets possible applications. Also, the optical system information is sometimes not completely defined as well as some of the datasets lack of sensor information, capture distance, range of spectrum in the generated images, and environmental validation. Some of the datasets only provide cropped image regions of the complete scene, so information like aperture, speed shutter, and sensitivity is lacking. When collecting with mobile devices, data from the IMU (i.e., an accelerometer and a gyroscope) may be beneficial in reducing the negative effects of the rolling shutter and recognizing motion blur (e.g., smartphones). In addition, several datasets only provide compressed images, reducing the quantity of data captured by the sensor.

Due to the differences in image quality, researchers require a complete explanation of the method and capture information in different research areas. Despite the common features in research problems, smartphone depth capture research focuses on using additional sensor information available in mobile platforms (IMU or multiple imaging sensors) and computational methods to process captured images, whereas depth in motion research focuses on novel sensors and optical systems.

Many research papers underline the absence of datasets suited for evaluating a specific parameter (i.e., a constrained environment with only one parameter’s variability), which leaves research conclusions and underlying reasons unclear, underlining the need for more research. In some cases, having a clear protocol description may be enough to solve the problem. If the camera specifications (usually removed for privacy concerns) were contained in the EXIF/metadata, several of these issues may be avoided. This information is generally missing from datasets created using custom-built cameras, as well as a protocol description. While many details of specialized hardware are hidden from users of other datasets, publicly accessible cameras provide such attributes by default in the image file.

There is also a mismatch between datasets acquired under visible light. In some cases, the authors used a monochromatic sensor with a band-pass filter to catch the entire visible band of light, while in others, they used mass market cameras to collect visible light in three spectral bands (separately for the colors red, green, and blue). Because the spectral sensitivity of the visible light filter differs from that of the individual color filters (even when the color bands are combined), they should not be compared. Additionally, most consumer color cameras have a Bayer filter that restricts individual band resolution to one-quarter for red and blue spectra and one-half for green; as a result, two-thirds of the color information are estimated rather than measured.

The review also found that synthetic image datasets have not got momentum in depth estimation research. Researchers prefer standard datasets (real) instead of synthetic images, despite the fact that synthetic images have a higher number of samples. The authors feel that these datasets lack the realism of research effects that occur in less confined circumstances.

Only a small percentage of distance depth capture research has focused on computational depth capture, such as using super-resolution, whereas the majority has focused on constructing a standard optical system with mirrors for the capture.

A. RELATED RESEARCH
This has been a review of existing datasets generated for performance evaluation, with a focus on depth. The datasets investigated in this work could be useful in other fields of research that use images of the human body, faces, poses, objects, indoor/outdoor, medical information, and environments.

Face tracking and segmentation have been used in a wide range of applications, from human–computer interaction to medical diagnosis. These applications usually have other well-known datasets, but they primarily share initial depth image processing, such as depth localization and segmentation. As a result, depth estimation datasets could be useful as a secondary data source. Furthermore, a useful medical diagnostic for detecting neurotransmitter and neuronal activity levels has been proven using the pupil [66]. Object recognition and classification algorithms are a comparable, but more sophisticated academic area. However, depth estimation is often a more difficult challenge. It’s been utilized in medical applications, such as diagnosing computer vision syndrome and facial recognition technologies.
Biometrics datasets are restricted in that they do not contain identification information, that restricts the use of many datasets. Alternatively, unsupervised methods can play an important role in depth-based recognition problems.

B. CHALLENGES AND COMPETITIONS

An independent evaluation and standard compression analysis can greatly help current depth estimation methods in a range of applications and tasks in computer vision research. There is a well-defined baseline for the SoA methods, but the results are greatly diverse due to the datasets, training, evaluation, and implementation methodologies. These variations make it difficult to compare the methods objectively for a specific problem related to depth estimation. Many of these issues can be avoided by creating benchmark datasets and conducting independent evaluations. This ensures an objective comparison of methods by using standardized protocols and environments. Competitions and/or challenges are commonly used to organize such evaluations. This strategy stimulates competition among academics in addition to the production of publicly available datasets with uniform measurements.

C. FUTURE RESEARCH DIRECTIONS

Image-based depth estimation using deep learning approaches has shown promising results following detailed research over the last few years. However, the subject is still in its early stages, and more developments are to be expected. In this section, the authors will go over some of the hot topics right now and point out in the right direction for future research.

• Data for training purposes is a problem: The availability of training data is critical to the effectiveness of deep learning algorithms. Unfortunately, compared to the training datasets used in tasks like classification and recognition, the size of publicly available datasets that comprise both images and their ground truth depth is small. Due to a lack of 3D training data, 2D supervision techniques have been utilized. However, many of them rely on silhouette-based supervision and can only reconstruct the visual hull as a result. Consequently, one can expect to see more papers in the future proposing new largescale datasets with diverse environments, new weakly-supervised and unsupervised methods that leverage various visual cues, and new domain adaptation techniques in which networks trained on data from a specific domain, such as synthetically rendered images, are adapted to a new domain, such as in-the-wild images, with very little retraining and supervision. Research into realistic rendering approaches that can bridge the gap between actual and synthetically created images has the potential to help with the training data problem.

• Generalization to unseen objects: Most SoA studies, such as BTS and AdaBins, divide a dataset into three subsets for training, validation, and testing, and then report on the performance on the test subsets. However, it is unclear how these approaches would perform on categories of objects/images that have never been seen before. In reality, the ultimate goal of the depth estimation method is to be able to recreate any 3D shape from any set of images. Learning-based strategies, on the other hand, only work on images and objects that are part of the training set. A number of recent publications have attempted to examine this topic. However, combining classical and learning-based strategies to improve the generalization of the latter methods would be an interesting direction for future research.

• Fine-scale depth estimation: The coarse depth structure of shapes can be recovered using current SoA approaches. Although subsequent work has enhanced the resolution of the reconstruction by employing refinement modules, thin and small portions such as plants, hair, eyes, and fur remain unrecoverable.

• Reconstruction versus recognition: The difficulty of obtaining depth from images is ill-posed. As a result, effective solutions must incorporate low-level image cues, structural knowledge, and a high-level understanding of the object. Deep learning-based depth estimation algorithms are biased towards recognition and retrieval, according to a recent study [8]. As a result, many of them have difficulty generalizing and recovering fine-scale features. Therefore, it is expected that this area of research might see more exploration in the future on how to mix top-down (i.e., recognition, classification, and retrieval) and bottom-up approaches (i.e., pixel-level reconstruction based on geometric and photometric cues). This has the potential to improve the approaches’ generalization capabilities (see item (2) above).

• Handling multiple objects in the presence of occlusions and cluttered backgrounds: Most of the SoA approaches deal with single-object images. Images taken in the wild, on the other hand, often feature a variety of things from several categories. Detection and reconstruction within regions of interest have been used in previous studies. The modules for detection, depth, and reconstruction are all independent of one another. These tasks, however, are interrelated and might benefit from one other if completed together. Two major concerns must be solved in order to achieve this goal. The first is a lack of multiple-object reconstruction training data. Second, especially for methods that are learned without 3D supervision, creating proper CNN architectures, loss functions, and learning procedures is critical. In general, these employ silhouette-based loss functions, which necessitate precise object segmentation.

• Data Imbalance: Some class representations are limited in some scene understanding tasks, such as semantic labelling, whereas others have a lot of examples. Learning a model that respects both types of categories and performs equally well on frequent and less frequent ones is a challenge that requires more research.

Deep-learning algorithms for depth estimation rely largely on training datasets annotated with ground truth labels, which are
difficult to come by in the actual world. Large datasets for 3D reconstruction are expected to emerge in the future. One of the interesting future paths for study in depth estimation is emerging new self-adoption algorithms that can adapt to changing circumstances in real-time or with minimal supervision.

IX. SUMMARY

This analysis reveals significant heterogeneity in available datasets in terms of size (ranging from 5 to >1,800 classes), sensors used, image quality, and so on. Because of this variation, there is a dataset available for many research issues, but it is not always straightforward for researchers to choose the optimal alternative. This analysis not only serves to help researchers find the right dataset and loss function, but it also makes suggestions for establishing new ones. Because there are so many features that researchers can be interested in, presenting a global summary in the form of a research article is challenging. According to the bibliometric analysis, the KITTI dataset is the most cited, followed by CITYSCAPES and NYU-V2 datasets. As a result, it is recommended that these datasets be used as benchmarks when comparing approaches to the published SoA. Furthermore, a license signed by a researcher is sufficient to get these datasets, as opposed to the signature of the institutional legal representative, which is normally requested by others. It’s best to use datasets developed for specific challenges or competitions for comparative research because they come with a standardized evaluation methodology. MOBILE-RGBD is a tool for evaluating depth images obtained by smartphone cameras. FACESCAGE is a framework for studying 3D reconstruction and detection. There are 360° and WEB STEREOSTEREO VIDEO to examine combinations of multiple modalities. [68] has put a lot of effort into developing publicly available datasets, in addition to KITTI and CITYSCAPES. Their website contains 102 high-quality datasets (plus more from other modalities), making it the most comprehensive web resource the authors found. Although the bibliometric analysis showed that these datasets are not as popular as those at KITTI or CITYSCAPES, NYU-V2 and did not cover the depth estimation-based research, it is encouraged that the academics explore them further.

X. RECOMMENDATION FOR BUILDING A COMPREHENSIVE DATASETS

Various scientific groups have explored important aspects of gathering and distributing research data.

- **Plan availability for years to come** - In the field of depth estimation, the acceptance of a new benchmark is typically difficult. It is critical to allocate resources for database distribution for several years into the future in order to maintain the database’s availability. The most important resources are (i) technical – a solid URL for the promoting website as well as the infrastructure to keep it available – and (ii) personal – a designated person responsible for licensing maintenance as well as answering any problems that prospective users may encounter.
  - Make access simple - We discovered that databases that include licenses that can be signed by individual academics are more popular. For young researchers, requiring the signature of the legal institutional representative, especially in a college environment (usually the rector), is a substantial barrier. Instead, they frequently choose to develop their own database. If an institutional representative’s signature is required, we recommend posting the whole license agreement as well as a sample of the database images on the project website. This aids in determining whether the database is appropriate for a certain research project before beginning the administrative procedures required to secure the requisite approvals.
  - Include a statistically relevant number of samples
  - Acquiring and handling test subjects is one of the most challenging tasks when creating a biometric database. The number of subjects included should be as large as possible; however, there is always a minimum size for obtaining statistically relevant results. Although this minimum is difficult to quantify for the general case, the statistical significance of 100 samples obtained from the same subjects is not the same as 1000 samples obtained from 100 different subjects.
  - Make the database unique - Many authors who use a database in one publication continue to use it in subsequent publications. A database is often used to investigate particular qualities or problems in a methodical manner, as we have seen in earlier sections. A successful database should assist users in coming up with new research findings and conclusions. As a result, the database should be able to meet the needs of new study areas where benchmarks have yet to be created. With this review, the authors hope to aid in this work by making the demands more apparent to database designers.
  - Extensive protocol and setup description - Despite the fact that the majority of the datasets available were developed to test a specific hypothesis or for a certain study aim, researchers frequently suggest that the dataset can be beneficial for more than one research topic. It is critical to offer a detailed description of the technique and setup in order to maximize the dataset’s potential. Important information, such as the wavelength of the setup lighting, the distance at which the images were captured, and descriptions of the sensor or optical system employed, is usually lacking, restricting the usability of the datasets.
  - More Challenging Datasets - For depth estimation and instance segmentation, several large-scale image datasets have been generated. However, new complex datasets, as well as datasets for diverse types of images,
are still needed. Datasets containing a large number of objects and overlapping objects would be quite useful for still images. This may make it possible to train models that are better at dealing with dense object scenarios and high overlaps between objects, which are typical in real life. With the growing popularity of 3D image depth reconstruction, particularly in autonomous vehicles and robotics, large-scale 3D image datasets are in high demand. The creation of these datasets is more difficult than that of their lower-dimensional equivalents. Existing datasets for 3D image depth estimation are often insufficiently large, and some are synthetic, therefore larger and more difficult 3D image datasets can be extremely beneficial.

XI. CONCLUSIONS

This paper provides a detailed review of the depth datasets and loss functions developed in the field of computer vision for depth estimation problems. The publicly available depth datasets and depth-based loss functions have achieved impressive performance in various depth map tasks based on deep learning networks. People detection and action recognition, faces and poses, perception-based navigation (i.e., street signs, roads), object and scene recognition, and medical applications are among the five general categories in which the depth datasets are categorized. Each depth dataset’s main properties and characteristics are described and compared. To generalize model results across different environments, a mixing approach for depth datasets is presented. In addition, depth estimation loss functions are briefly presented, which will facilitate in the training of deep learning depth estimation models on a variety of datasets for both short- and long-range depth map estimation. Three of the most popular datasets are evaluated using SoA deep learning-based depth estimation algorithms. Finally, there is a discussion of challenges and future research, as well as recommendations for creating comprehensive depth datasets, which will help researchers in choosing relevant datasets and loss functions for evaluating their results and methods.

The main aim of this survey paper is that, to speed up the research in depth estimation tasks and compare the results to SoA methodologies for use case applications, researchers in this discipline must first understand the appropriate depth datasets and loss functions. To improve generalization, researchers should incorporate various datasets during training, validation, and testing. However, when combining datasets with different features, caution is required. The network’s design and building blocks are important, but its performance is mostly influenced by how it is trained, which requires a diverse dataset and an appropriate loss function.

REFERENCES

[1] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in CVPR 2011, pp. 1521–1528.
[2] C. Wang, S. Lucey, F. Perazzi, and O. Wang, “Web stereo video supervision for depth prediction from dynamic scenes,” in 2019 International Conference on 3D Vision (3DV), 2019, pp. 348–357.
[3] W. Chen, Z. Fu, D. Yang, and J. Deng, “Single-image depth perception in the wild,” Adv. Neural Inf. Process. Syst., vol. 29, pp. 730–738, 2016.
[4] T. Schops et al., “A multi-view stereo benchmark with high-resolution images and multi-camera videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3260–3269.
[5] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, “Tanks and temples: Benchmarking large-scale scene reconstruction,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–13, 2017.
[6] H. Sarbolandi, D. Lefloch, and A. Kolb, “Kinect range sensing: Structured-light versus Time-of-Flight in Kinect,” Comput. Vis. Image Underst., vol. 139, pp. 1–20, 2015.
[7] F. Khan, S. Hussain, S. Basak, J. Lemley, and P. Corcoran, “An efficient encoder–decoder model for portrait depth estimation from single images trained on pixel-accurate synthetic data,” Neural Networks, vol. 142, pp. 479–491, 2021.
[8] X.-F. Han, H. Laga, and M. Bennamoun, “Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 5, pp. 1578–1604, 2019.
[9] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image segmentation using deep learning: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., 2021.
[10] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer,” IEEE Trans. Pattern Anal. Mach. Intell., p. 1, 2020, doi: 10.1109/TPAMI.2020.3019967.
[11] F. Khan, S. Salahuddin, and H. Javidnia, “Deep learning-based monocular depth estimation methods—A state-of-the-art review,” Sensors, vol. 20, no. 8, p. 2272, 2020.
[12] A. Bhoi, “Monocular depth estimation: A survey,” arXiv Prepr. arXiv:1901.09402, 2019.
[13] C. Zhao, Q. Sun, C. Zhang, Y. Tan, and F. Qian, “Monocular depth estimation based on deep learning: An overview,” Sci. China Technol. Sci., pp. 1–16, 2020.
[14] Y. Ming, X. Meng, C. Fan, and H. Yu, “Deep Learning for Monocular Depth Estimation: A Review,” Neurocomputing, 2021.
[15] A. Mertan, D. J. Duff, and G. Unal, “Single Image Depth Estimation: An Overview,” arXiv Prepr. arXiv:2104.06456, 2021.
[16] H. Song, J. Hong, H. Choi, and J. Min, “Concrete Delamination Depth Estimation Using a Noncontact MEMS Ultrasonic Sensor Array and an Optimization Approach,” Appl. Sci., vol. 11, no. 2, p. 592, 2021.
[17] J. K. Devine, E. D. Chino, R. R. Markwald, L. P. Schwartz, and S. R. Hursh, “Validation of Zulu Watch against Polysomnography and Actigraphy for On-Wrist Sleep-Wake Determination and Sleep-Depth Estimation,” Sensors, vol. 21, no. 1, p. 76, 2021.
[18] P. Liu, Z. Zhang, Z. Meng, and N. Gao, “Monocular Depth Estimation with Joint Attention Feature Distillation and Wavelet-Based Loss Function,” Sensors, vol. 21, no. 1, p. 54, 2021.
[19] K. PNVR, H. Zhou, and D. Jacobs, “Sharingan: Combining synthetic and real data for unsupervised geometry estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13974–13983.
[20] J. Spencer, R. Bowden, and S. Hadfield, “Defeat-Net: general monocular depth via simultaneous unsupervised representation learning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14402–14413.
[21] H. Javidnia and P. Corcoran, “Accurate depth map estimation from small motions,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 2455–2461.
[22] S. Bazarbakan, H. Javidnia, and J. Lemley, “Semiparallel deep neural network hybrid architecture: first application on depth from monocular camera,” J. Electron. Imaging, vol. 27, no. 4, p. 1, 2018, doi: 10.1117/1.jei.27.4.043041.
[23] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, “On the
uncertainty of self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3227–3237.

K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, “Structure-guided ranking loss for single image depth prediction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 611–620.

V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Aigand, “3d packing for self-supervised monocular depth estimation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2485–2494.

H. Javidnia and P. Corcoran, “Real-time automotive street-scene mapping through fusion of improved stereo depth and fast feature detection algorithms,” in 2017 IEEE International Conference on Consumer Electronics (ICCE), 2017, pp. 225–228.

L. Spinello and K. O. Arras, “People detection in RGB-D data,” in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2011, pp. 3838–3843.

E. Cippitelli, E. Gambi, S. Gasparini, and S. Spinase, “TST fail detection dataset v2,” IEEE DataPort. IEEE. https://doi.org/10.1109/HEVC/7C, 2014.

Z. Li et al., “Learning the depths of moving people by watching frozen people,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4521–4530.

F. Olfi, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, “Berkeley mhad: A comprehensive multimodal human action database,” in 2013 IEEE Workshop on Applications of Computer Vision (WACV), 2013, pp. 53–60.

B. Kwolek and M. Kepkpi, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” Comput. Methods Programs Biomed., vol. 117, no. 3, pp. 489–501, 2014.

D. Vaufreydaz and A. Ngre, “Mobilerbd, an open benchmark corpus for mobile rgb-d related algorithms,” in 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), 2014, pp. 1668–1703.

G. Fanelli, T. Weise, J. Gall, and L. Van Gool, “Real time head pose estimation from consumer depth cameras,” in Joint pattern recognition symposium, 2011, pp. 101–110.

R. Min, N. Kose, and J. L. Dugelay, “KinectfaceDB: A Kinect database for face recognition,” IEEE Trans. Syst. Man, Cybern. Syst., vol. 44, no. 11, pp. 1534–1548, 2014, doi: 10.1109/TSMC.2014.2331215.

G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, “POSEidon: Face-from-Depth for driver pose estimation,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017, vol. 2017-Janua, pp. 5494–5503, 2017, doi: 10.1109/CVPR.2017.583.

H. Yang et al., “Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 601–610.

E. Nesli and S. Marcel, “Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect,” in IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS’13), 2013, pp. 1–8.

A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3354–3361, 2012, doi: 10.1109/CVPR.2012.6248074.

I. M. Alvarez, T. Gevers, Y. LeCun, and A. M. Lopez, “Road scene segmentation from a single image,” in 2012 Conference on Computer Vision, 2012, pp. 376–389.

R. Zhang, S. A. Candra, K. Vetter, and A. Zakhor, “Sensor fusion for semantic segmentation of urban scenes,” in 2015 IEEE international conference on robotics and automation (ICRA), 2015, pp. 1850–1857.

G. Ros, S. Ramos, M. Granados, A. Bakhtiari, D. Vazquez, and A. M. Lopez, “Vision-based offline-online perception paradigm for autonomous driving,” in 2015 IEEE Winter Conference on Applications of Computer Vision, 2015, pp. 231–238.

M. Cordts et al., “The cityscapes dataset for semantic urban scene understanding,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, “Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 899–908.

J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, “Sparsity invariant cnns,” in 2017 international conference on 3D Vision (3DV), 2017, pp. 11–20.

Z. Bauer, F. Gomez-Donoso, E. Cruz, S. Orts-Escalon, and M. Cazorla, “AUASL, a large-scale high-resolution outdoor stereo dataset,” Sci. data, vol. 6, no. 1, pp. 1–14, 2019.

J. Hidalgo-Carrió, D. Gelhrig, and D. Scaramuzza, “Learning monocular dense depth from stereo events,” in 2020 International Conference on 3D Vision (3DV), 2020, pp. 534–542.

A. Sharma and J. Ventura, “Unsupervised learning of depth and ego-motion from cylindrical panoramic video,” in 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), 2019, pp. 58–58.

N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in European conference on computer vision, 2012, pp. 746–760.

I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, “Joint 2d-3d semantic data for indoor scene understanding,” arXiv Prepr. arXiv1702.01105, 2017.

J. Xiao, A. Owens, and A. Torralba, “Sun3d: A database of big spaces reconstructed using sfm and object labels,” in Proceedings of the IEEE international conference on computer vision, 2013, pp. 1625–1632.

S. Song, S. P. Lichtenberg, and J. Xiao, “Sun rgb-d: A rgb-d scene understanding benchmark suite,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 567–576.

Z. Li and N. Snavely, “MegaDepth: Learning Single-View Depth Prediction from Internet Photos,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2041–2050, 2018, doi: 10.1109/CVPR.2018.00218.

I. Vasiljevic et al., “Diodo: A dense indoor and outdoor depth dataset,” arXiv Prepr. arXiv1908.00463, 2019.

D. Scharstein et al., “High-resolution stereo datasets with subpixel-accurate ground truth,” in German conference on pattern recognition, 2014, pp. 31–42.

H.-A. Le, T. Mensink, P. Das, S. Karazoglou, and T. Gevers, “Eden: Multimodal synthetic dataset of enclosed garden scenes,” in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1579–1589.

S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, “Semantic scene completion from a single depth image,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1746–1754.

A. Chang et al., “Matterport3d: Learning from rgb-d data in indoor environments,” arXiv Prepr. arXiv1709.06158, 2017.

A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, “Taskonomy: Disentangling task transfer learning,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3712–3722.

Q.-H. Pham, M. A. Uy, B.-S. Hua, D. T. Nguyen, G. Roig, and S.-K. Yeung, “Lcd: Learned cross-domain descriptors for 2d-3d matching,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2020, vol. 34, no. 07, pp. 11856–11864.

N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, “Omnidepth: Dense depth estimation for indoors spherical panoramas,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 448–465.

D. Wang, “MineNav: An Expandable Synthetic Dataset Based on Minecraft for Aircraft Visual Navigation,” arXiv Prepr. arXiv2008.0854z, 2020.

A. Saxena, M. Sun, and A. Y. Ng, “Make3d: Learning 3d scene structure from a single still image,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 5, pp. 824–840, 2008.

J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, “A benchmark for the evaluation of RGB-D SLAM systems,” in
2012 IEEE/RSI international conference on intelligent robots and systems, 2012, pp. 573–580.

K. B. Ozoruk et al., “EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos,” Med. Image Anal., vol. 71, p. 102058, 2021.

V. Srivastav, T. Issenhuth, A. Kakkathodamohammadi, M. de Mathelin, A. Gangi, and N. Padov, “MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation,” arXiv Prepr. arXiv.1808.08180, 2018.

A. P. Twintada, S. Shehata, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padov, “Endonet: a deep architecture for recognition tasks on laparoscopic videos,” IEEE Trans. Med. Imaging, vol. 36, no. 1, pp. 86–97, 2016.

N. L. Rodas, F. Barrera, and N. Padov, “See it with your own eyes: Markerless mobile augmented reality for radiation awareness in the hybrid room,” IEEE Trans. Biomed. Eng., vol. 64, no. 2, pp. 429–440, 2016.

B. Unmenhofer et al., “Demon: Depth and motion network for learning monocular stereo,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5038–5047.

J. M. Facil, B. Unmenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, “CAM-Convs: Camera-aware multi-scale convolutions for single-view depth,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11826–11835.

O. Sener and V. Koltun, “Multi-task learning as multi-objective optimization,” arXiv Prepr. arXiv:1810.04650, 2018.

S. Niklaus, L. Ma, J. Yang, and F. Liu, “3D ken burns effect from a single image,” ACM Trans. Graph., vol. 38, no. 6, pp. 1–15, 2019.

S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4009–4018.

G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci, “Transformers solve the limited receptive field for monocular depth prediction,” arXiv Prepr. arXiv:2103.12091, 2021.

J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, “From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation,” 2019, [Online]. Available: http://arxiv.org/abs/1907.10326.

J. Chang and G. Wetzstein, “Deep optics for monocular depth estimation and 3D object detection,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10193–10202.

M. Song, S. Lim, and W. Kim, “Monocular Depth Estimation Using Laplacian Pyramid-Based Depth Residuals,” IEEE Trans. Circuits Syst. Video Technol., 2021.

R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188.

H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep Ordinal Regression Network for Monocular Depth Estimation,” Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 2002–2011, 2018, doi: 10.1109/CVPR.2018.00214.

A. Tao, K. Sapra, and B. Catanzaro, “Hierarchical multi-scale attention for semantic segmentation,” arXiv Prepr. arXiv:2005.10821, 2020.

Y. Yuan, X. Chen, X. Chen, and J. Wang, “Segmentation transformer: Object-contextual representations for semantic segmentation,” in European Conference on Computer Vision (ECCV), 2021, vol. 1.

R. Mohan and A. Valada, “Efficientps: Efficient panoptic segmentation,” Int. J. Comput. Vis., vol. 129, no. 5, pp. 1551–1579, 2021.

B. Cheng et al., “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12475–12485.

X. Zhang et al., “Denas: Densely connected neural architecture search for semantic image segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13956–13967.

J.-H. Lee and C.-S. Kim, “Multi-loss rebalancing algorithm for monocular depth estimation,” in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX. 2020, pp. 785–801.

J. Y. Jason, A. W. Harley, and K. G. Derpanis, “Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness,” in European Conference on Computer Vision, 2016, pp. 3–10.

Faisal Khan received the BS in Mathematics and M.Sc in Computer Science from University of Peshawar, Pakistan in 2002 and 2005, respectively. He received MS and PhD degrees in Computer Engineering from Jeonbuk National University, South Korea in 2016 and 2020, respectively. He was awarded with Jeonbuk National University presidential award for academic excellence during his PhD studies. He had worked as a postdoctoral researcher at Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. He is currently working as a postdoctoral research fellow at National University of Ireland, Galway, Ireland. His research interests include smart grid, energy management, electric vehicles, optimization algorithms, micro-grid operations, distributed energy resources, Peer-to-peer energy trading, and image processing using fuzzy logic, game theory, ontologies, AI, and block-chain approaches and technologies.

Shahid Hussain received BS in Mathematics and M.Sc in Computer Science from University of Peshawar, Pakistan in 2002 and 2005, respectively. He received MS and PhD degrees in Computer Engineering from Jeonbuk National University, South Korea in 2016 and 2020, respectively. He was awarded with Jeonbuk National University presidential award for academic excellence during his PhD studies. He had worked as a postdoctoral researcher at Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. He is currently working as a postdoctoral research fellow at National University of Ireland, Galway, Ireland. His research interests include smart grid, energy management, electric vehicles, optimization algorithms, micro-grid operations, distributed energy resources, Peer-to-peer energy trading, and image processing using fuzzy logic, game theory, ontologies, AI, and block-chain approaches and technologies.

Shubhajit Basak received the B.Tech. degree in Electronics and Communication Engineering from the West Bengal University of Technology, India, in 2011 and M.Sc. in Computer Science from the National University of Ireland Galway, Ireland in 2018. He had more than 6 years of industrial experience as a software development professional. He is currently pursuing the Ph.D. degree in Computer Science with the National University of Ireland Galway, Ireland. He is also with FotoNation/Xperi. His research interest includes deep learning tasks related to computer vision.

Mohamed Moustafa (Member, IEEE) received his B.Sc. (Honours) in Computer Science and Information Technology from the National University of Ireland, Galway, graduating with first-class honours in 2021. During his undergraduate studies he was awarded the University Scholar title three years in a row by the university. He is currently pursuing a Ph.D. degree in Electrical and Electronics Engineering at the National University of Ireland, Galway as part of his employment-based postgraduate programme jointly funded by
the Irish Research Council and the Xperi Corporation, where he is currently employed. His research interests include computer vision, deep learning, embedded systems, edge-AI, and their applications for health monitoring.

PETER CORCORAN (Fellow, IEEE) holds the Personal Chair in electronic engineering at the College of Science and Engineering, National University of Ireland Galway. He is currently an IEEE Fellow recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He was a Co-Founder in several start-up companies, notably FotoNation, now the Imaging Division of Xperi Corporation. He has over 600 technical publications and patents, over 100 peer-reviewed journal articles, 120 international conference papers, and a co-inventor of more than 300 granted U.S. patents. He is a member of the IEEE Consumer Electronics Society for over 25 years. He is the Editor-in-Chief and the Founding Editor of IEEE Consumer Electronics Magazine.