electronics

MDPI

Review

A Survey on Deep Learning Based Methods and Datasets for Monocular 3D Object Detection

Seong-heum Kim [1] and Youngbae Hwang [2,*]

[1] Department of Smart Systems Software, Soongsil University, Seoul 06978, Korea; seongheum@ssu.ac.kr
[2] Department of Electronics Engineering, Chungbuk National University, Cheongju 28644, Korea
* Correspondence: ybhwang@cbnu.ac.kr; Tel.: +82-43-261-3641

Abstract: Owing to recent advancements in deep learning methods and relevant databases, it is becoming increasingly easier to recognize 3D objects using only RGB images from single viewpoints. This study investigates the major breakthroughs and current progress in deep learning-based monocular 3D object detection. For relatively low-cost data acquisition systems without depth sensors or cameras at multiple viewpoints, we first consider existing databases with 2D RGB photos and their relevant attributes. Based on this simple sensor modality for practical applications, deep learning-based monocular 3D object detection methods that overcome significant research challenges are categorized and summarized. We present the key concepts and detailed descriptions of representative single-stage and multiple-stage detection solutions. In addition, we discuss the effectiveness of the detection models on their baseline benchmarks. Finally, we explore several directions for future research on monocular 3D object detection.

Keywords: deep learning; monocular object detection; 3D object detection

Citation: Kim, S.-H.; Hwang, Y. A Survey on Deep Learning Based Methods and Datasets for Monocular 3D Object Detection. *Electronics* **2021**, *10*, 517. https://doi.org/10.3390/electronics10040517

Academic Editor: Rashid Mehmood

Received: 28 December 2020
Accepted: 11 February 2021
Published: 22 February 2021

1. Introduction

Deep learning networks have increasingly been extending the generality of object detectors. In contrast to traditional methods in which each stage is individually hand-crafted and optimized by classical pipelines, deep learning networks achieve superior performance by automatically deriving each stage for feature representation and detection. In addition, new approaches for data-driven representation and end-to-end learning with a substantial number of images have led to significant performance improvements in 3D object detection. With the evolution of deep representation, object detection is being widely used in robotic manipulation, autonomous driving vehicles, augmented reality, and many other applications, such as CCTV systems.

Beyond the significant progress in image-based 2D object detection, 3D understanding of real-world objects is an open challenge that has not been explored extensively thus far. In addition to the most closely related studies [1–6], we focus on investigating deep learning-based monocular 3D object detection methods. For location-sensitive applications, conventional 2D detection systems have a critical limitation in that they do not provide physically correct metric information on objects in 3D space. Hence, 3D object detection is an interesting topic in both academia and industry, as it can provide relevant solutions that significantly improve existing 2D-based applications.

Camera sensors that capture color and texture information have emerged as an essential imaging modality in many computer vision applications. The passive camera sensors do not interfere with other active optical systems, and always work well with them when needed. For image-based deep representations that encode depth cues, monocular images are also highly cost-effective. Owing to considerable accumulations of annotations for RGB databases, the data-driven representations using deep neural networks make monocular 3D object detectors even more advantageous without expensive depth-aware sensors or cameras at additional viewpoints.

To understand the major breakthroughs and current progress in practical 3D object detection, we contribute to the literature by reviewing recent developments in deep learning-based state-of-the-art 3D object detection with monocular RGB databases. The remainder of this paper is organized as follows. Section 2 presents the overall background for our taxonomic approach. Section 3 summarizes well-known datasets for monocular 3D object detection. Section 4 comprehensively describes multi-stage approaches and end-to-end learning for monocular 3D object detection methods. The key concepts, representative solutions, and effectiveness of the detection models in terms of their baseline benchmarks are discussed in detail. Section 5 briefly highlights potential research opportunities. Finally, Section 6 concludes the paper.

2. Background on Object Detection

Given an image with a pixel grid representation, object detection is the task of localizing instances of objects with bounding boxes of a certain class. An important contribution in solving the 2D object detection problem is the use of region-based convolutional neural networks (R-CNNs), which involves two main stages: region proposal and detection. The region of interest (ROI) of an image is proposed on the basis of certain assumptions, such as color, texture, and size. The ROI is cropped to feed a CNN that performs the detection. By combining prior knowledge and labeled datasets, the two-stage detection framework has emerged as a classical model in both 2D and 3D object detection [7–10].

Another important algorithm for object detection is the YOLO algorithm [11]. It does not have a separate region proposal stage; instead, it divides an input image roughly into an N × N grid. Based on each grid cell, localization and classification tasks are performed together in a unified regression network, followed by further post-processing. Early end-to-end approaches performed poorly in the detection of small or occluded objects. As new datasets are being developed, there have been significant innovations in end-to-end networks [12–14]. As fewer proposal steps with hand-crafted features are involved in single-stage methods, they are computationally less complex than multi-state approaches that usually prioritize detection accuracy. In practice, there was active competition between multi-stage and single-stage methods for object detection tasks. 3D object detection is similar to this overall flow.

The goal of 3D object detection systems is to provide 3D-oriented bounding boxes for 3D objects in the 3D real world. The 3D cuboids can be parameterized by 8-corners, 3D centers with offsets, 4-corner-2-height representations, or other encoding methods. In monocular 3D object detection methods, we seek the oriented bounding boxes of 3D objects from single RGB images. Similarly to 2D-image-based object detection systems, monocular 3D object detection methods can be also categorized into two main types, as shown in Figure 1. From a taxonomic point of view, we have extended them to six sub-categories, according to the main distinguishing features of each sub-category. As shown in Table 1, we have summarized the main features of ten high-quality datasets, such as descriptions with quick links, input data types, contextual information for different applications, the availability of synthetic RGB images, the number of 3D object instances/categories, the number of training/testing images, and lastly, other related references, which can be used for future research. In Table 2, we have briefly explained key features of the most representative works for each category and the related databases, computational time, and so on. All of those methods use powerful algorithms that can only run on a high-performance system using GPUs, and we did not pay attention to lightweight deep learning models for lower-power embedded/mobile systems.

Based on a general understanding of object detection, we review 11 datasets for monocular 3D object detection and more than 29 recent algorithms. The unique properties of 3D object detection systems, such as different data representations and the availability of both 2D and 3D annotations, make the 3D detection frameworks more complicated and interesting.

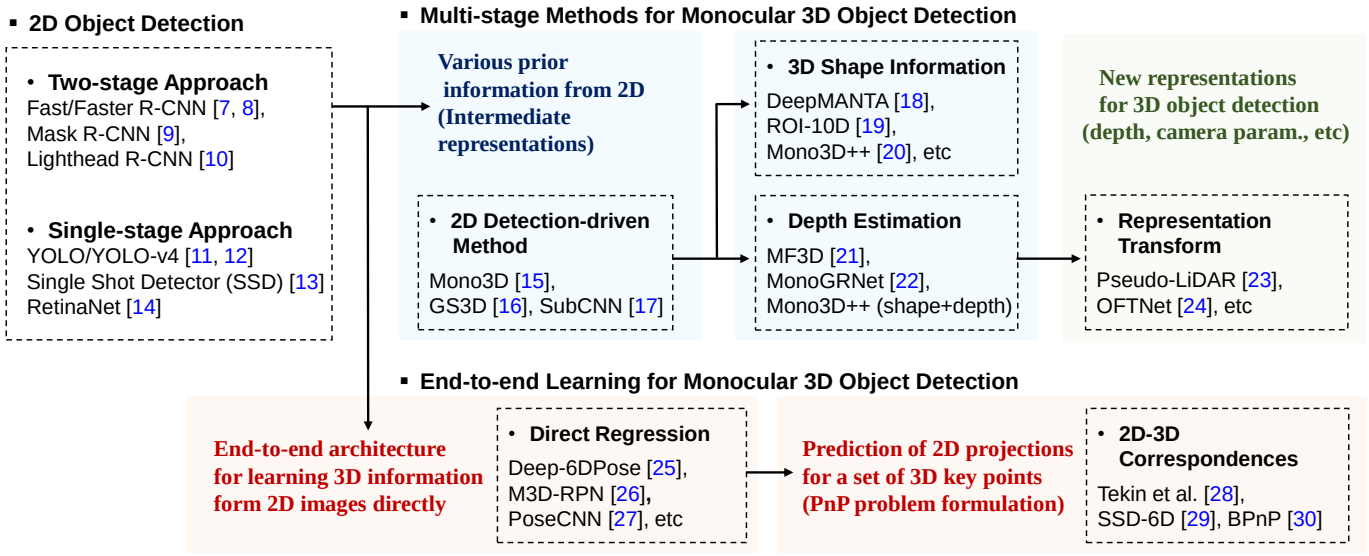

Figure 1. Overview of the 2D/3D monocular object detection methods [7–30].

Table 1. Datasets used for monocular 3D object detection.

Dataset	Description	Data Type	Scene Type	Syn.?	# 3D Objects	# Images	Related References
PASCAL 3D+ [31]	A Benchmark for 3D Object Detection in the Wild (WACV 2014)	RGB + 3D models	Indoor + Outdoor	Real	3000 per cate. 12 categories	>20,000	PASCAL VOC [32], ImageNet [33], Google Warehouse
SUN RGB-D [34]	A Scene Understanding Benchmark Suite (CVPR 2015)	RGB-D	Indoor	Real	14.2 per image 800 categories	>10,335 in total	NYU depth v2 [35], BerkeleyB3DO [36] SUN3D [37]
ObjectNet 3D [38]	A Large Scale Database 3D Object Recognition (ECCV 2016)	RGB + 3D models	Indoor + Outdoor	Real + Syn.	201,888 inst. 100 objects	90,127	ImageNet [33], ShapeNet [39], Trimble Warehouse
FAT [40]	A Synthetic Dataset for 3D Object Detection (CVPRW 2018)	RGB + 3D models	Household Objects	Syn.	1–10 per image 21 objects	61,500	YCB [41]
BOP [42,43]	Benchmark for 6D Object Pose Estimation (ECCV 2018, ECCVW 2020)	RGB-D + 3D models	Indoor (various)	Real + Syn.	302,791 inst. in 97,818 real images (test) 171 objects (w/texture)	>800 K train, test RGB-D (mostly synthetic)	LM [44], LM-O [45], T-LESS [46], ITODD [47], YCB-V [27], HB [48], RU-APC [49], IC-BIN [50], IC-MI [51], TUD-L [42], TYO-L [42]
Objectron [52]	Object-Centric Videos in the Wild with Pose Annotations	RGB	Indoor + Outdoor	Real	17,095 inst. (multi-view) 9 categories	>4 M (14,819 videos)	Open Images [53] Similar to CAMERA [54] (Real/syn. data)
KITTI 3D [55]	KITTI Vision Benchmark Suite—3D Objects (CVPR 2012)	RGB (Stereo) + PointCloud	Driving Scenes	Real	80,256 inst. 3 categories	14,999	Virtual KITTI 2 [56] * Photo-realistically simulated DB
CityScape 3D [57]	Dataset and Benchmark for 9 DoF Vehicle Detection (CVPRW 2020)	RGB (Stereo)	Driving Scenes	Real	8 categories	5000	CityScape [58]
Synscapes [59]	A Photo Synthetic Dataset for Street Scenes	RGB	Driving Scenes	Syn.	8 categories	25,000	Similar to CityScape [58] (Structure, content)
SYNTHIA-AL [60]	Synthetic Collection of Imagery and Annotation —3D Boxes (CVPR 2012)	RGB	Driving Scenes	Syn.	3 categories	>143 K	ImageNet [33] SYNTHIA [61]

Table 2. Representative monocular 3D object detection methods.

Method	Category	Key Feature	Related Datasets	Computational Time	Code?
Mono3D [15]	2D-driven Method	An energy minimization approach that places object candidates on the 3D plane, and then scores each candidate box via several intuitive potentials encoding semantic segmentation, contextual information, size and location priors and typical object shape.	KITTI 3D	It takes 1.8 s in a single core, but exhaustive search in the proposal step can be done efficiently as all features can be computed with integral images.	Yes
DeepMANTA [18]	3D Shape Informat.	Simultaneous vehicle detection, part localization (even if some parts are hidden), visibility characterization, and 3D template for each detection. Coarse-to-fine object proposal with multiple refinement steps for accurate 2D vehicle bounding boxes.	KITTI 3D	It is approximately twice faster than Mono3D, due to the lower resolution of images in the coarse-to-fine method, considerably reducing a search space.	No
MF3D [21]	Depth Estimation	Multi-level fusion scheme for monocular 3D object detection utilizing a standalone depth estimation module to ensure the accurate 3D localization and improve the detection performance.	Cityscape, KITTI 3D	The inference time including the depth module achieves about 120 ms per img. on a NVIDIA GeForce GTX Titan X.	No Partial implement.
Pseudo-LiDAR [23]	Represent. Transform	Conversion of an estimated depth map from stereo or monocular imagery into a 3D point cloud, which mimics the real LiDAR, and takes advantage of existing LiDAR-based detection pipelines.	KITTI 3D	The paper does not focus on real-time processing. More effective way to speed up depth estimation is required.	Yes
Deep-6D Pose [25]	Direct Regression	An end-to-end deep learning framework for detection, segmentation, and 6D pose estimation of 3D objects. It directly regress 6D object poses without any post-refinements.	LineMOD (LM), [51]	Due to the end-to-end architecture, it offers an inference speed of 10 fps on a Titan X GPU (not optimized speed).	No Partial implement.
Tekin et al. [28]	2D-3D Correspo.	A single-shot approach for simultaneously detecting an object in an RGB image and predicting its 6D pose without requiring multiple stages or having to examine multiple hypotheses. It predicts the projected vertices of the object's 3D bounding box.	LM, LM-O	A pose refinement step can be used to boost the accuracy, but it runs at 10 fps. Without additional post-processing, it takes 50 fps on a single Titan X GPU.	Yes

3. Datasets Used for Monocular 3D Object Detection

Although deep learning methods for 2D object detection using pure RGB images have achieved considerable success, it is much more challenging to obtain 3D-oriented bounding boxes owing to the absence of absolute 3D information in the 2D image plane. In general, when the number of layers to be trained increases, the size of the labeled datasets is especially important for obtaining the data-driven solution. Compared with well-built 2D datasets, 3D datasets are still under construction. In this section, we review well-known RGB (or RGB-D) datasets used in recent 3D object detection tasks.

3.1. Beyond PASCAL

PASCAL3D+ [31], which is an extension of one of the most popular 2D detection benchmarks, PASCAL VOC [32], handles 12 selected categories of rigid objects. As shown in Figure 2, 3D CAD models are collected and aligned to images in the PASCAL VOC database. To overcome some ambiguities of 2D images in different categories, additional photos from ImageNet [33], according to the 12 categories, are included. Instead of a small number of images per category, captured in controlled environments, more than 3000 objects per category are stored in PASCAL3D+, with rich 3D annotations for objects appearing in a variety of natural images. Indeed, PASCAL3D+ with its extended 3D information, facilitated significant progress in research on monocular 3D object detection.

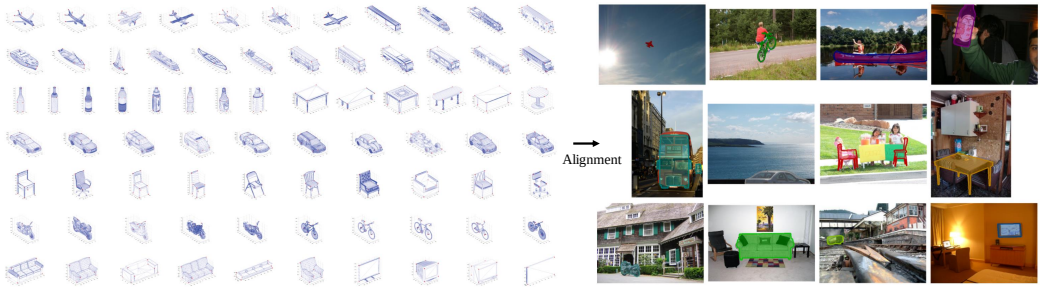

Figure 2. PASCAL3D+ [31] for 3D object detection and pose estimation.

3.2. SUN RGB-D

SUN RGB-D [34], which is an extension of the SUN3D dataset [37] developed at Princeton University, contains 10,355 images with depth channels from four different sensors. For example, 3389 frames without severe motion blur have been manually selected from the SUN3D videos. Further, 1449 RGB-D images from the NYU Depth V2 dataset [35] and 554 realistic scene images from the Berkeley B3DO dataset [36] are included. The collected datasets handle 47 scene categories and around 800 object categories, and the annotations consist of 146,617 2D polygons and 64,595 3D-oriented bounding boxes. On average, 14.2 objects are annotated in each image. Thus, the SUN RGB-D dataset has had a major impact on indoor vision tasks such as 3D object, detection using RGB or RGB-D images, object orientation estimation, and indoor scene understanding.

3.3. ObjectNet3D

ObjectNet3D [38] comprises 90,127 images with 44,147 3D models in 100 rigid object categories. The 2D images are initially acquired from ImageNet [33] and added from Google searches for some categories that do not include sufficient numbers of images. Furthermore, 3D models from ShapeNet [39] and Trimble Warehouse were selected to precisely align with 201,888 objects appearing in these photos. Similarly to PASCAL3D+, this process gives a 3D shape label and the closest pose annotation for each object. Given accurate 2D and 3D annotations, ObjectNet3D facilities the study of object proposal, shape retrieval, object detection, and pose estimation algorithms.

3.4. Falling Things (FAT)

Falling Things (FAT) [40], which is an extension of the Yale-CMU-Berkeley (YCB) dataset [41], contains 61,500 snapshots of 21 household objects. A physics-based graphic simulator was introduced to generate photorealistic training images and automatic annotations to evaluate and train robotic manipulation algorithms for household scenes. By combining synthetic objects and backgrounds, all the information, such as 2D/3D locations, poses, and segmentation masks, is available for all the objects drawn in the high-quality simulation images. The simulation process and analysis are also well described. In the context of robust perception for robotic manipulation, this synthetic dataset can help improve the overall performances of object classification algorithms, pose recognition algorithms, and other related algorithms.

3.5. Benchmark for 6D Object Pose Estimation (BOP)

The Benchmark for 6D Object Pose Estimation (BOP) dataset [42,43] contains training images with rigid objects at various viewpoints, wherein the 6D poses (3D translation and 3D rotation in space) of the presented objects are known, or texture-mapped models of the 3D objects were well prepared. The images with test objects have occlusions or background clutter; hence, some parts of the objects may not be observable and only the visible surface can fit multiple 3D models. For the evaluation, the benchmark additionally consists of eight well-known datasets in different scenarios.

One of the datasets is the LineMOD (LM) benchmark [44], comprising 15 texture-less objects with discriminative shapes, sizes, and colors in household environments. A test

image with background clutter shows an annotated object with small occlusions only. The level of occlusion is further controlled in the LineMOD-Occluded (LM-O) dataset [45] with additional annotations of all associated objects. T-LESS [46] comprises 30 industry-relevant objects from 20 scenes with discriminative colors and no significant textures. The objects have mutual similarities and symmetries in size and shape, and some objects are composited from other assemblable objects. The MVTec Industrial 3D Object Detection Dataset (ITODD) [47] contains 3500 labeled scenes and 28 objects acquired from realistic setups for industrial applications. The 6D poses are known for the validation images only and are not available publicly for the test images. The YCB-Video (YCB-V) dataset [27] contains 133,827 frames with 21 objects, selected from 92 videos of the YCB dataset. The 80 K simulation images in the original dataset are also included in this benchmark. In the case of the HomebrewedDB (HB) dataset [48], there are 33 toy, household, and industry-relevant objects in 13 complex scenes with different backgrounds.

As shown in Figure 3, other datasets, such as RU-APC [49], IC-BIN [50], IC-MI [50], and TYO-L [42], were also used for the BOP Challenge. The training and test images for 3D object detection are annotated with ground-truth object poses. Every dataset, together with the given 3D models, is available in the unified BOP format.

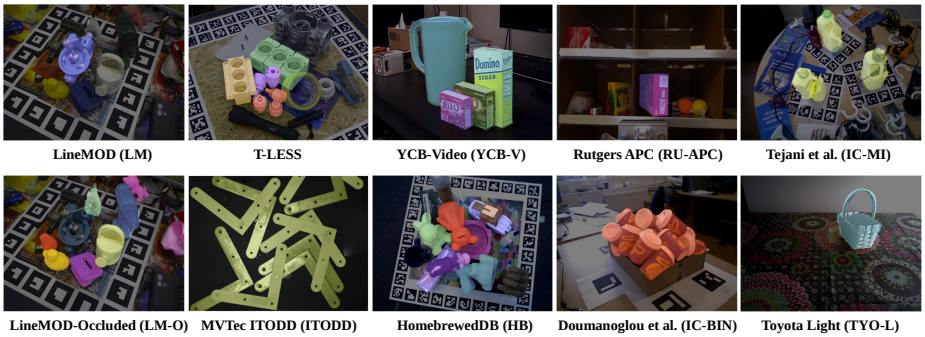

Figure 3. Benchmark for 6D Object Pose Estimatio (BOP) [43]. Note that multiple datasets for 3D object detection were also collected for evaluation.

3.6. Context-Aware MixEd ReAlity (CAMERA)

The Context-Aware MixEd ReAlity (CAMERA) dataset [54] addresses the limitations of traditional data generation by synthetically generating a large number of training images and ground truths in a faster and more cost-effective manner. From mostly tabletop scenes, 553 background images were acquired in widely varying conditions. For hand-scale objects such as a bowl, a bottle, a can, a camera, a mug, and a laptop, selected from ShapeNetCore, a total of 300,000 composited images of 31 indoor tabletop scenes were rendered. Further, 25,000 photorealistic images were set aside for validation. For point sampling and plane detection, the mixed reality compositing technique exploits the Unity engine with custom plugins. In contrast to non-context-aware images from previous approaches, the simulated images in this database facilitate and improve generalization in learning-based methods.

3.7. Objectron

The Objectron dataset [52] is a collection of short, object-centric video clips that are accompanied by AR session metadata, including camera poses, sparse point clouds, and characterizations of planar surfaces in the surrounding environment. In each video, the camera moves around the object, capturing it from different angles. The data also include manually annotated 3D bounding boxes for each object, which describe the object's position, orientation, and dimensions. The dataset consists of 15,000 annotated video clips supplemented with over 4 million annotated images of the following objects: bikes, books, bottles, cameras, cereal boxes, chairs, cups, laptops, and shoes. In addition, to ensure geo-diversity, our dataset was collected from 10 countries across five continents. Along with the dataset, we must mention a 3D object detection solution for four categories of

objects: shoes, chairs, mugs, and cameras. These models were trained using this dataset and released in MediaPipe, Google's open-source framework for cross-platform customizable ML solutions for live and streaming media.

3.8. KITTI 3D

The KITTI3D benchmark [55] comprises 7481 training images, no official validation images, and 7518 test images. As there is no validation set, the training images are often split into 3712 images for training and 3769 images for analyzing the validation results before reporting the results on the test set via the evaluation server. For 3D annotations of 2D images, the possible 3D bounding boxes are given for only three categories, namely, cyclist, car, and pedestrian. Depending on object truncation, occlusion, and distance to the camera, the difficulty of 3D detection is determined as hard, moderate, or easy. Figure 4 shows examples of the target objects with their ground-truth bounding boxes.

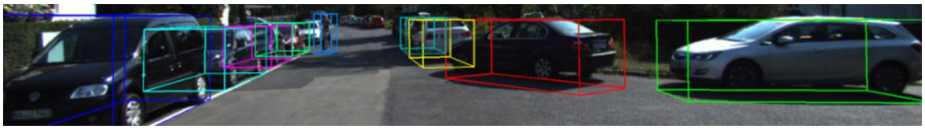

Figure 4. KITTI 3D object detection benchmark [55].

Virtual KITTI [56] is one of the first synthetic datasets for training and testing machine learning models for autonomous driving applications. In a video-game world, it is easy to create data for rare events, and scenes with changes in only one condition (such as the weather) can be generated. Moreover, the exact ground truth can be generated along with simulated images; hence, little annotation is required. The Unity game engine has been used to explore this concept by carefully recreating real-world videos from the popular KITTI autonomous driving benchmark suite.

3.9. CityScape 3D

Cityscapes 3D [39] is an extension of Cityscapes [57], one of the most influential datasets, which enriches annotations with high-quality 3D bounding boxes for vehicles. The original Cityscapes dataset [58] contains 5000 images, of which 2975 are used for training, 500 are used for validation, and 1525 are used for testing. The 3D bounding box annotations cover all eight semantic classes in the vehicle category of the Cityscapes dataset, i.e., bicycle, bus, car, caravan, train, motorcycle, trailer, and truck. The 3D annotations are newly labeled with nine degrees of freedom (DoF) using stereo images, resulting in more accurate re-projection in images and a higher range than LiDAR-based methods. It is a new benchmark for 3D detection tasks in autonomous driving with full 3D orientation, including yaw, pitch, and roll labels. Compared to other 3D detection datasets, Cityscapes 3D has a high object density, which indicates complex scenes.

3.10. Synscapes

The Synscapes database [59], created by a collaboration between 7DLabs Inc. and researchers at Linköping University, is a synthetic dataset comprising more than 25,000 simulated images. In the context of street scene parsing, the photorealistic rendering technique tries to capture every aspect of the optical process in the camera system, from illumination sources such as the sun, to the object's material and geometric composition, and finally, to the sensors. As photons hit digital sensors through a lens in a pinhole camera, the signal is converted into an image with other physically plausible noise. For example, owing to the relative velocities of vehicles, motion blur can be modeled. Synscapes ensures simulations that are representative of the real world for data augmentation in driving scenes.

3.11. SYNTHetic Collection of Imagery and Annotations (SYNTHIA-AL)

The SYNTHetic collection of Imagery and Annotations (SYNTHIA) dataset [61] provides photorealistically rendered frames in city-level scenes. The categories handled in the database are building, bicycle, car, fence, lane marking, road, pedestrian, pole, sidewalk, sky, traffic light, traffic sign, vegetation, and void. The SYNTHIA-AL dataset is generated by modifying the SYNTHIA environment using the Unity Pro game engine. In the context of driving scenarios, the data are generated in a virtual world consisting of three different areas, namely, town, city, and highway. These areas are populated with a variety of pedestrians, cars, cyclists, and wheelchairs, except for the highway, which is limited to cars. Several environmental conditions, such as season (winter, fall, and spring), day time (day or night), and weather (clear or rainy) can be set. The ground truth is provided in terms of 2D/3D bounding boxes, instance segmentation, and depth information [60].

4. Monocular 3D Object Detection Methods

Researchers have proposed new methods to overcome challenges for monocular 3D object detection. Here, we categorize these methods into multi-stage and end-to-end approaches.

4.1. Multi-Stage Approaches

First of all, we can deal with an ill-posed problem by employing prior hypotheses on 3D objects. The prior knowledge includes semantic, context, shape, and location information, and so on. By performing distinct tasks linearly, including hand crafting features, 2D boxes of interest can be proposed. Alternatively, we can use standard 2D object detectors with simple deep neural networks. As an example of GS3D [16], 2D detections are converted into basic 3D boxes using projection knowledge, which is called guidance. Given the guidance, the 3D bounding boxes are further refined without expensive stereo data or point clouds.

With an additional 3D shape prior, we can perform 3D object detection through CAD template matching. During the detection process, a template library will be established, and the network will match the best model in the template library. In the case of the method in [19], the 3D template, partial visibility, and partial coordinates of the detected vehicle are given. Then, these features are considered to estimate the localization and orientation with 2D–3D model fitting. Even if some parts of the test objects are not visible in the 2D case, vehicle models can be retrieved via template matching.

As monocular images lack depth information owing to the principle of perspective transformation, we can use deep learning to predict the depth map of the image first, which serves as the basis for 3D object detection in the next stage. To achieve effective monocular depth estimation, many algorithms have been developed in recent years. In addition to using the depth estimation module, the object ROI and depth feature map are fused to calculate the object coordinate and spatial location information [21]. Using a multi-layer fusion scheme, this framework [21] can generate the final pseudo-point cloud information for its application.

Likewise, it is also a popular algorithm for converting the image information into point cloud information; the point-cloud-related network is then used for processing. For another application of representation transform, the orthographic feature transform (OFT) [24] maps perspective images to orthographic bird's eye view (BEV) images in the deep-learning-based framework. In general, the representation transform selects an application-specific data representation that is more suitable for the target scenario than the image domain. Hence, it can achieve satisfactory detection results.

4.1.1. 2D Detection-Driven Methods

Based on PASCAL3D+ [31], simultaneous 2D object detection and viewpoint estimation was proposed by Su et al. [62]. Given an input RGB image and a bounding box from an off-the-shelf detector, a deep representation was tailored specifically for viewpoint estimation. The authors selected category-specific orientations of objects with a novel

loss layer adapted over synthetically generated viewpoint labels. Experimental results indicated that the performances of both joint detection and viewpoint estimation can be significantly improved on PASCAL 3D+. 2D images and 3D shapes/scans can be connected through their image synthesis pipeline; thus, information can be transported between the two domains bidirectionally. When training datasets for deep learning need to be manually annotated, this approach infers 3D information with negligible human effort.

For encoding raw images with different sensor modalities in compact descriptors, Wohlhart et al. [63] used pair-wise and triplet-wise constraints on training images and template views. By considering the dissimilarity and similarity of the descriptors, they efficiently captured both the object categories and the 3D poses. The constraints untangle the input images with different objects from different views into several clusters, which are not only well separated but also structured as the corresponding sets of 3D poses. The Euclidean metric between descriptors is sufficiently large when the descriptors are given from different objects. Furthermore, when the encoded descriptors are given from the same object, the distance is directly associated with the different 3D poses of the object. In this manner, the learned descriptors can be generalized to classify unseen objects as well. This approach requires binary masks of the objects of interest; however, it works well with either RGB or RGB-D images of the LineMOD dataset [44].

To use a standard CNN method for high-quality detections, Chen et al. [15] assumed that objects are always on the ground plane. Initially, given a set of category-specific object proposals, the monocular 3D object detection is formulated as an energy minimization task that optimally locates object candidates in the 3D world. Based on prior information such as object size, location, shape, segmentation, and contextual information, each intuitive loss function accurately optimizes a 3D box. Hence, Mono3D [15] uses two stages with a 2D object detection network. The detection performance of this approach was quantitatively confirmed on the challenging KITTI benchmark.

When objects have significant truncation, occlusion, and scale variations in the CNN-based detection pipeline, region proposals can often be a bottleneck. To alleviate this issue, subcategory-aware CNNs [17] have an interesting region proposal network whereby the proposal step is guided by subcategory information. The subcategory concept refers to categories of objects that share similar attributes, such as 3D shape and pose. Based on this key assumption, the SubCNN considers a new detection network for joint detection and subcategory classification. In addition, test objects at different scales are handled using image pyramids in an efficient manner. While exploring the effects of subcategory information on CNN-based object detection, extensive experiments were conducted on the PASCAL VOC 2007, PASCAL3D+, and KITTI detection benchmarks.

In GS3D [16], 2D detections are converted into basic 3D boxes using projection knowledge, which is called guidance. Given the guidance, the 3D bounding boxes are further refined without expensive stereo data or point clouds. To remove representation ambiguities in 2D bounding boxes, the underlying 3D structure in the surface feature is extracted. In practice, coarse cuboids are reported to have sufficient accuracy for determining the 3D bounding boxes of objects by refinement. To refine the 3D detections, the surface feature extraction module, which is an affine extension of RoIAlign, is also used. In this framework, the complex residual regression problem is reformulated as a classification task, which is much easier to train. Finally, the discriminative ability is enhanced by the quality-aware loss. This approach was evaluated using the KITTI 3D benchmark.

Figure 5 shows 2D detection-driven GS3D. 2D detections and the orientations of target objects are obtained using the CNN-based model (2D+O subnet). Then, the proposed algorithm generates the guidance using the given 2D bounding box and orientation with a projection matrix. The refinement model (3D subnet) uses the extracted features from visible parts and 2D detections of the projection guidance. Instead of direct regression, the reformulated classification task is adopted by the refinement model with the quality-aware loss to achieve a more accurate result.

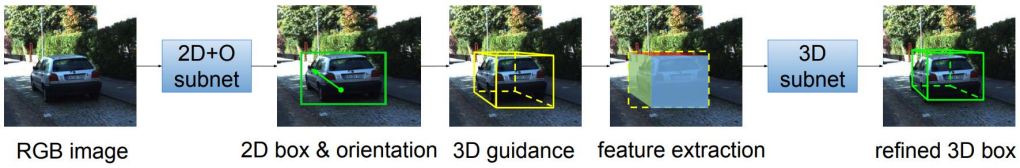

Figure 5. Overview of GS3D [16].

4.1.2. 3D Shape Information

While Mono3D [15], an optimization-based pioneering method, does not show satisfactory accuracy and speed, its successor, Mono3D++ [20], achieves improved performance with better template matching, as does the Ceres toolbox [64]. Mono3D++ [20] uses coarse and fine 3D hypotheses to infer the object shape and pose from one RGB image. Specifically, a fine representation for vehicles is generated by morphable wireframe models with different shapes and poses. For lower sensitivity to 2D landmark features, a coarse representation aims to model 3D bounding boxes to improve stability and robustness. For joint energy minimization with a projection error, three priors are considered, namely, vehicle shape, a ground plane constraint, and unsupervised monocular depth.

3D shape information-based methods tend to become slow when the number of shape templates or object poses increases, because hand-crafted steps for comparing them are required for optimization. To tackle this problem with some physical quantities, Konishi et al. [65] proposed a new image feature based on orientation histograms of random projection images from CAD models. Similarly, in [66], coarse initialization was adopted for 3D poses of texture-less objects. In [67], temporally consistent, local color histograms were used for pose estimation and segmentation of rigid 3D objects. For handheld objects, the statistical descriptors can be learnt online within a few seconds.

Instead of optimizing separate quantities, Chabot et al. [18] proposed a multi-tasking network structure for 2D and 3D vehicle analysis from a single image. For simultaneous part localization, visibility characterization, vehicle detection, and 3D dimension estimation, the many-tasks network (MANTA) first detects 2D bounding boxes of vehicles in multiple refinement stages. For each detection, it also gives the 3D shape template, part visibility, and part coordinates of the detected vehicle even if some parts are not visible. Then, these features are considered to estimate the vehicle localization and orientation using 2D–3D correspondence matching. To access the 3D information of the test objects, the vehicle models are searched for template matching. The real-time pose and orientation estimation uses the outputs of the network in the inference stage. At the time of publication, this approach was the state-of-the-art approach using the KITTI 3D benchmark in terms of vehicle detection, 3D localization, and orientation estimation tasks.

As shown in Figure 6, an input image is passed forward to the deep MANTA network where convolution layers with the same weights have the same color. The existing architecture is split into three blocks. With these networks, the object proposals are refined iteratively until the final detection that is associated with the part's coordinate, the part's visibility, and the template similarity. Moreover, non-maximum suppression (NMS) removes some redundant detections. Based on the outputs, the best 3D shape is chosen in the inference stage. 2D and 3D pose computation is then performed with the associated shape.

In the ROI-10D algorithm [19], a monocular deep network directly optimizes a novel 3D loss formulation and then lifts a 2D bounding box to 3D shape recovery and pose estimation. Using CAD templates and synthetic data augmentation, deep feature maps are generated and combined to obtain the shape dimensions. Then, shape regression is performed to obtain the object information. In particular, the pose distributions are well analyzed in the KITTI 3D benchmark. In metrically accurate pose estimation, learning synthetic data is useful for increasing the pose recall; however, some hand-crafted modules such as 2D and 3D NMS have a strong influence on the final results.

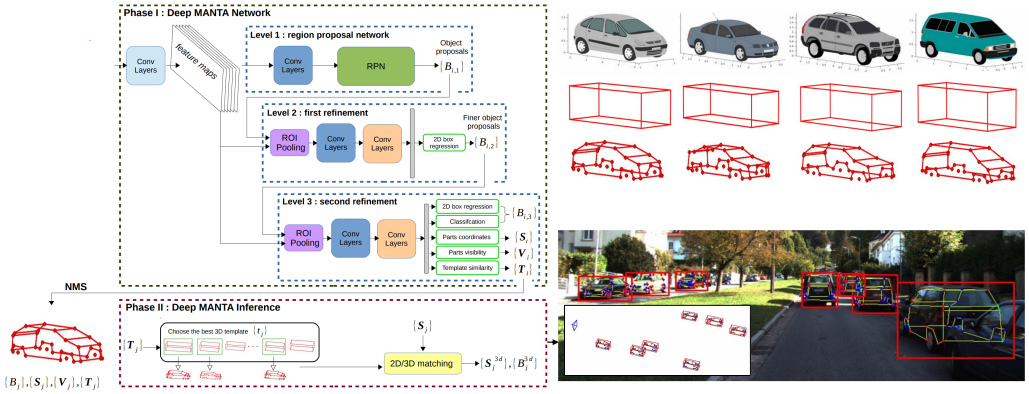

Figure 6. Overview of the deep MANTA approach [18].

4.1.3. Depth Estimation

On the basis of deep-learning-based monocular depth estimation, Xu and Chen [21] proposed the multi-level fusion-based 3D object detection (MF3D) algorithm, which combines the Deep3Dbox algorithm [68] and a standard depth estimation module. Using deep CNN features, it basically uses the existing detectors. In addition to 2D proposals, the disparity estimation is computed to generate a 3D point cloud. Thus, the deep features derived from the RGB images and the point cloud are fused to enhance the object detection performance. The depth feature map and the ROI for objects are combined to obtain the 3D spatial location. The key idea of this framework is to use the multi-level fusion scheme, taking advantage of the standalone module for disparity computation. Experimental results showed that the performance of 3D object detection can be boosted by 3D localization. Figure 7 shows sub-networks of the MF3D framework [21]. The task-specific modules are responsible for objectness classification, 2D box regression, and disparity prediction. Based on region proposals and point cloud maps from the estimated disparity, the 3D bounding box of the object is optimized and visualized as shown in the figures on the right.

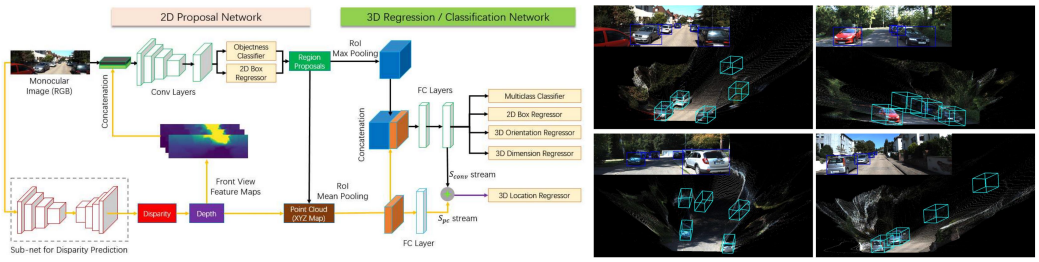

Figure 7. MF3D for 3D object detection from monocular images [21].

MonoGRNet [22] uses depth estimation for similar reasons. However, it does not require precise pixel-level depth annotation; only instance-level depth is considered for 3D localization. The unified MonoGRNet uses four sub-networks for instance depth estimation (IDE), 2D object detection, 3D localization, and corner regression. In the IDE module, the depth of a target object is predicted at the center of its 3D bounding box. With sparse supervision, this network performs depth inference only on the areas of objects detected as 2D bounding boxes. By avoiding depth estimation for the entire image, it reduces the computational requirements considerably. The global 3D position is achieved by simply estimating the object location in the vertical and horizontal dimensions. Then, the corner coordinates are regressed in the local context. By optimizing the poses and positions of 3D bounding boxes, MonoGRNet is trained in the global context.

4.1.4. Representation Transform

Pseudo-LiDAR [23] assumes that the main innovation for bridging the gap between LiDAR-based and pure image-based 3D object detection is the computational representation itself for expressing the 3D scene. In other words, the point cloud representation may be more suitable for monocular 3D object detection than the image-based representation with the same quality of depth information. Thus, the image-based depth maps are converted into the proposed pseudo-LiDAR representations mimicking real liDAR signals. For this reason, the deep ordinal regression network (DORN) [69] has been exploited for monocular depth estimation, and the mathematical relationship between the 2D image coordinates and the 3D pseudo-point cloud has been derived. After processing two additional networks for pseudo-point data, the representation transform makes it possible to apply any LiDAR-based algorithms to monocular 3D object detection.

The BEV image is another popular representation for applications involving autonomous vehicles. A common technique for converting images into BEV images is inverse perspective mapping (IPM); however, it typically assumes that all the pixels should be on the ground plane and it requires accurate camera parameters for estimating the plane homography. Without needing to calibrate extrinsic parameters, the OFT [24] maps perspective images to orthographic BEV images in the deep-learning-based framework. The overall architecture and its output is shown in Figure 8. To encode camera images, the authors used a front-end ResNet-18 architecture and accumulated image-based features into a voxel-based representation. The voxel features were then collapsed along the vertical dimension to yield orthographic ground plane features. Another network was finally employed to remove the distortional effects of perspective projection and refine the BEV map. The top-down network processes these features in the BEV space. At each location on the ground plane, it predicts a confidence score S, a position offset Δpos, a dimension offset Δdim, and an angle vector Δang. Reasoning in the 3D space improves the performance, and the network is robust to objects that are distant or occluded.

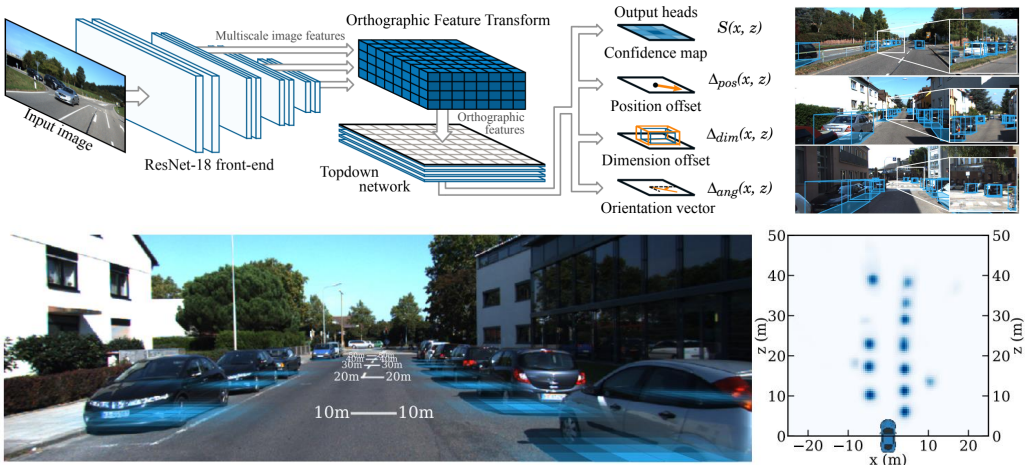

Figure 8. Architecture overview for the orthographic feature transform [24].

In fact, the method proposed in [70] is similar. The two-phase approach basically uses IPM to infer the distance from the 3D scene. In the first phase, camera motions such as pitch and roll rotations are removed using inertial measurement units. The front view is corrected and projected onto the BEV via the IPM module. In the second phase, the position, orientation, and size of the vehicle are detected by the CNN. By canceling the camera pitch and roll rotations, a vanishing point is moved to infinity so that it is not affected by any vehicle attitude. The resulting projection image is parallel and linear with respect to the x-y coordinate system of the vehicle. For 3D localization of objects in the real world, the bounding box detected from the BEV is transformed by the inverse projection matrix

for conversion into metric units. The proposed algorithm was quantitatively validated using KITTI 3D.

The representation transform is also a promising candidate for robotics, augmented reality, and 3D scene understanding. Wang et al. [54] recently proposed a novel normalized object coordinate space (NOCS) for indoor applications. It defines a shared space with consistent object orientation and scaling. To estimate the metrically accurate size and pose of unseen objects, the NOCS map is predicted by the proposed network and used with the depth map for pose fitting. Extensive experiments on the CAMERA dataset [54] demonstrated that the proposed method can estimate the sizes and poses of unseen object instances robustly in real environments.

4.2. End-to-End Approaches

Some recent methods directly return 3D location information of objects and pose parameters of a camera. For example, a well-known deep representation uses the shared 2D and 3D detection space to build an independent monocular 3D area recommendation network, which achieved the best performance at the time of its publication [26]. In practice, the space for directly searching for rotation parameters is nonlinear, making it difficult for the CNN to recognize the rotation of an object. To avoid this problem, some algorithms have been proposed to discretize the rotation space or refine the result iteratively. Post-processing is often crucial for direct regression.

Meanwhile, algorithms that use key points for an algebraic solver do not directly obtain the pose of the object from a monocular image, but focus on 2D–3D point correspondences to find a firm geometric model using the perspective-n-point (PnP) algorithm [71]. 2D key point detection is easier than 3D localization and rotation estimation; however, it requires a model of a known 3D object and some predefined key points. One of the clear trends is to increase the number of matching pairs. For example, a pixel-wise voting network (PVNet) [72] predicts pixel-level indicators corresponding to the key points so that they can handle truncation or occlusion of object parts. Each pixel votes for the predefined key points, which are optimized by the Ceres toolbox [64]. PVNet can achieve good results compared to the previous algorithms.

In this section, we will review these methods using end-to-end CNNs with monocular images.

4.2.1. Direct Regression

Mousavian et al. [68] proposed Deep3Dbox, a method that estimates the 3D pose and the size of the 3D bounding box of an object. Similarly to previous 2D-based object detectors, it partitions the parameter space of the 3D bounding boxes into multiple bins (MultiBin). From the shared convolution features, the proposed architecture estimates the dimensions, angles, and confidences using fully connected layers, which can facilitate robust MultiBin-based regression. Instead of using the L2 loss function to extract a rotation angle directly, the angle is separated into numerous bins. Then, the confidence of each bin and the offset are predicted using the residual of the center bin. In the object space size estimation, the L2 loss function is directly used to compute the offset of the space size. After determining the size and rotation angle of the object, we can restore the object's 6D pose by computing the rotation matrix of the object. This method outperforms other previous methods in terms of the orientation accuracy on the KITTI dataset and viewpoint estimation on Pascal3D+.

Xiang et al. [27] proposed PoseCNN for 6D object pose estimation. It consists of feature extraction, embedding, and classification/regression blocks. The feature extraction network is based on a single-shot detector (SSD) [13]. Here, the extracted features are shared among all the tasks performed by the second stage. Semantic labeling can provide rich information for the objects, and this pixel-level classification is effective at dealing with occlusions. Inspired by the implicit shape model, it can regress the center position and object distance. It is difficult for the CNN to regress the 3D rotation matrix directly owing

to the nonlinearity of the target space. Hence, a discretization scheme was proposed for the space of rotation. However, the accuracy of estimating the rotation matrix may be degraded by converting the regression of the rotation into a classification problem. To overcome this problem, in PoseCNN [27], two new loss functions were designed for estimating the 3D rotation matrix to handle symmetric objects and to match object shapes by decoupling the estimations of 3D translation and 3D rotation. Poirson et al. [73] also used SSD [13], the 2D object detector, to integrate the pose estimation for each detected object in the same network. The previous two-stage approach requires at least three resamplings of the image for region proposals, object detection, and pose estimation. By combining these steps into a single network, they achieved very fast object detection and pose estimation of up to 46 frames on a Titan X GPU.

Deep-6DPose [25] achieves simultaneous estimations of object detection, segmentation, and pose estimation using an end-to-end network. Interestingly, it takes advantage of the concept in Mask-RCNN [9], in order to directly regress 6D poses of objects without further post-processing. The remarkable contribution of this approach is the separate regression of translation and rotation matrices using a Lie group. Compared with a conventional orthonormal matrix or quaternion-based representation, a Lie algebra gives an optimal solution owing to fewer parameters and the unconstrained condition. Deep-6Dpose achieves rapid processing through its end-to-end architecture, and it is suitable for various robotic applications.

M3D-RPN [26] uses a shared space of 2D detection and 3D detection for a single 3D region proposal architecture. It gives greater weight to the relationship of 2D and 3D aspects. To improve the 3D parameter estimation accuracy, depth-aware convolution has been proposed for learning more high-level features with spatial information, as shown in Figure 9. Then, the pose optimization algorithm is adopted for orientation estimation, followed by 2D detections and 3D projections. Applying M3D-RPN to BEV and 2D and 3D object detection tasks shows the effectiveness of the single-stage network.

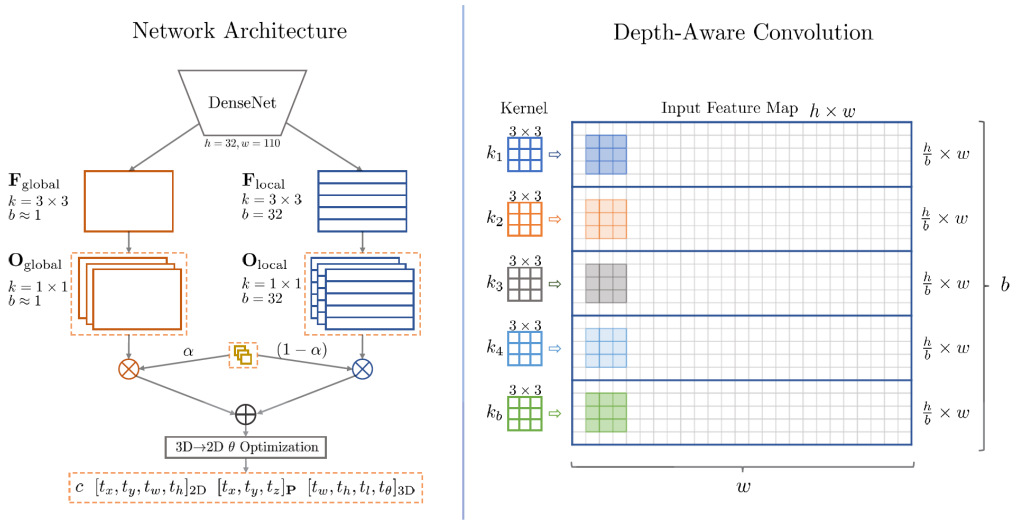

Figure 9. Overview of M3D-RPN [26]. It consists of parallel paths for global and local feature extraction.

Liu et al. [74] proposed measuring the degree of visual fitting between the object and the projected 3D proposals for achieving high-precision localization. After regressing the 3D bounding box and the orientation of the object for constructing suitable 3D proposals, they proposed the fitting quality network (FQNet), which can predict intersection over union (IoU) in the 3D space between the 3D bounding box and the target object using 2D cues, as shown in Figure 10. Their motivation was that denoting the projections on the image domain can provide additional knowledge to better understand the spatial relationship. Matching the object-rendered image with the input image generates better results compared to the limited accuracy of direct regression. DeepIM [75] is a new

refinement method that uses a deep neural network for matching 6D poses iteratively. Given an initial pose, a relative transformation can be predicted by matching the rendered image with the observed image. As rendering the object and estimating the 6D pose are complementary, the accuracy of pose estimation increases with iteration. The separate representation of 3D position and rotation not only achieves accurate estimated poses but also allows unseen objects to be refined. Experiments on commonly used benchmarks such as LM [44] and T-LESS [46] demonstrated that the proposed method shows significant improvements over previous methods.

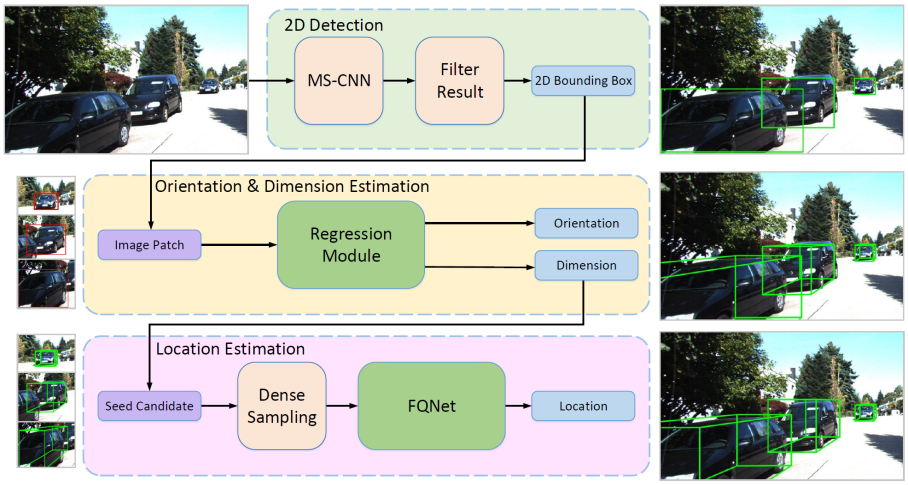

Figure 10. Overall pipeline of a deep fitting degree scoring network [74], which refines an initial bounding box using a regression module and FQNet.

4.2.2. 2D–3D Correspondences

BB8 [76] is based on the idea of using 2D–3D correspondences for 3D object localization. In the first step, a network of object segmentation is applied to an input image for localizing the objects. Next, another network is used to estimate the 2D projections of the interest points of the 3D boundaries around the target objects. The 6D pose is estimated using the relationship between the 3D bounding box corners and the corresponding projected 2D points. To handle the rotational symmetry, it restricts the pose ranges in the training stage and introduces a classifier to estimate the pose ranges at run time. For the final refinement of the estimated poses, it includes a feedback loop to compare the input image and the rendered object for better prediction of the 2D projected points. This holistic approach showed more accurate results on the challenging T-LESS dataset [46].

SSD-6D [29] applies the SSD concept [13] to 6D pose estimation. As an extension of SSD for inferring the 3D location and orientation, it predicts the corner points in the bounding boxes, classes, viewpoints, and in-plane rotation. For better results, it tries to find a proper sampling for the space of rotation. Interestingly, SSD-6D is trained using only a synthetic dataset, which can alleviate the difficulties of building a database for new target objects. SSD-6D can treat depth as an optional modality for hypothesis verification and pose refinement.

Tekin et al. [28] used the YOLO network [77] to predict the key points of corresponding objects. The network has a regular grid to present feature maps spatially. In each cell, the 2D positions of the corner points corresponding to the 3D bounding boxes are predicted. Then, the 6D pose can be computed using the PnP algorithm for the given 2D and 3D correspondences. However, the predicted 2D points may be insufficient for pose estimation when there is severe occlusion. To overcome such problems due to occlusion or truncation, Hu et al. [78] proposed an image-segmentation-based method to estimate an object's 6D pose by aggregating numerous local pose estimates, which can achieve more accurate key point estimations, even in cases of severe occlusion. To combine the pose candidates into a more robust set of 3D and projected 2D correspondences, confidence measures are

computed. Even in the case of severe occlusion, because it generates precise results based on merging local pose estimates robustly, it does not adopt an additional refinement process. The proposed algorithm was tested on the challenging LM-O [45] and YCB-V [27] datasets. We believe the future goal of these 2D–3D correspondence approaches is to incorporate the PnP step into the network to establish a complete, end-to-end framework.

The dense pose object detector (DPOD) [79] predicts dense multi-class 2D and 3D correspondence maps between input images and possible 3D models. A 6D pose is computed on the basis of the PnP algorithm and RANdom SAmple Consensus (RANSAC) for the correspondences. Then, the pose is refined from the initial pose estimation using the refinement architecture. In contrast to the previous methods that regress projections of the object's bounding boxes [28,76] or formulate pose estimation as a discrete classification problem [29], DPOD shows more robust and accurate 6D pose estimation owing to the dense correspondences. PVNet [72] also uses a denser key point prediction method, as shown in Figure 11. Instead of using sparse key points by regression or prediction, PVNet is used to predict pixel-level indicators corresponding to the key points. This flexible representation can handle occlusion or truncated key points robustly. The RANSAC-based voting scheme provides the spatial probability distribution of each key point for estimating 6D poses with an uncertainty-driven PnP algorithm.

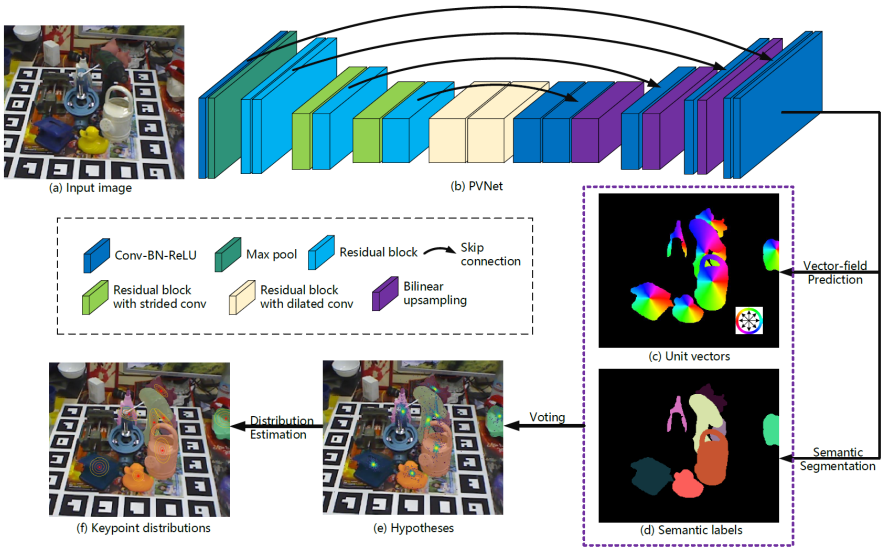

Figure 11. Overview of the keypoint localization in PVNet [72]. The probability distributions of the keypoint locations are estimated from hypotheses.

As the pose estimation problem belongs to the domain of geometric vision, it is essential to approach it as an end-to-end optimization to seamlessly combine the geometrically relevant information with the deep learning process. To this end, the BPnP [30] was proposed as an effective network module that computes the gradients of backpropagation by guiding parameter updates in the network using a PnP solver. If the optimization block is differentiable, the gradients of the PnP solver can be derived accurately via implicit differentiation. Although it integrates a layer from the PnP solver, the proposed method can be effectively employed to learn feature representations for various geometric vision problems such as structure from motion, geometric camera calibration, and pose estimation. For pose estimation, a BPnP-based trainable pipeline achieves higher accuracy by incorporating the feature map loss with 2D–3D reprojection errors.

5. Discussion

In images captured by cameras, geometric clues are essentially lost during dimension reduction through 3D to 2D projection. To overcome this problem, the studies reviewed in this paper constitute an active domain of 3D object detection using RGB images. To begin

with, we summarized well-known benchmark datasets [31,34,38,40–42,54,55,57,59,61] built by research groups from academia and industry. Benchmark databases are useful for fair comparison of the previous methods in the fields of machine learning and computer vision. They freely present high-quality training and test datasets. This is important for most of the deep-learning-based problems; in particular, for 3D-related tasks, sufficiently comprehensive information with a large amount of data is necessary. In fact, 3D bounding box annotation requires more specific guidelines for annotators and considerable time and effort.

When no single annotation alone is sufficient for ideal end-to-end training, we often use different types of human annotations together for a new task. By exploiting easily accessible 2D/3D databases, multi-stage methods typically have intermediate representations or features learnt from different annotations/guidance. The 2D detection-driven methods [15–17,62,63] started with two-stage 2D detectors, and developed the feature representation of RGB images for detecting 3D objects. The 3D information of objects is often inferred through fusion schemes with hand-crafted features on points, patches, and parts, or topological structures in 2D.

For the additional prior knowledge, 3D hypothesis has long been used to recognize and localize a 3D object from a single RGB image. To represent objects reliably, edges or more robust local features are extracted from a photo and matched with their counterparts in 3D models. Being that they are conceptually similar to the traditional approaches, the algorithms using a 3D shape hypothesis [18–20,65–67] can derive 3D information based on template matching with known 3D CAD models. Although it is not easy to deal with multi-object cases in real time, this approach can be highly practical, providing new deep representations and efficient optimization. Recognizing object locations in the actual 3D space also plays an important role in scene perception.

By inferring the scene geometry from 2D images, the depth information can compensate for the weakness of monocular vision. Given only a single RGB image and sufficient ground-truth depth data on the Web, we can predict the depth value of each pixel of the object of interest using learning-based monocular depth estimation. For example, DORN [69] is a popular network for depth extraction that incorporates multi-scale features to estimate pixel-level depth information with small errors. On the basis of existing depth extraction networks, many 3D object algorithms [20–22] combine such depth information as a sub-block in their proposed networks.

Some researchers argue that monocular 3D object detection is difficult to infer in perspective image-based representation, especially when the appearance and scale of objects vary drastically with depth and meaningful distances. A typical approach for the representation adaption is to transform the 2D image into a 3D point cloud. Then, we can use the available networks for processing the 3D point cloud. Some studies such as [23] have suggested that the data format for point cloud data is more suitable for detecting and recognizing 3D objects. Another way to alleviate occlusion and scale variation in perspective views is to convert the images into orthographic BEV images [24,70]. This approach forms the basis for future exploration of other tasks where the BEV representation is naturally applicable, such as 3D object tracking and motion forecasting. To address the representation challenge for hand-scale objects on a plane, Wang et al. [54] approached it as a problem of detecting correspondences in the normalized coordinates of a shared space of object description.

The most recent trend in monocular 3D object detection is learning deep neural networks to directly regress the 6D pose from a single image [25–27,68,75] or to estimate the 2D positions of 3D key points and solve the PnP algorithm [28–30,72,76,78,79]. The efficient, robust PnP algorithm can detect multiple 3D objects from the candidate correspondences between 2D and 3D points, but the object is considered as a global body in such cases. Consequently, these methods suffer from severe occlusion, and they easily fail in various real-world situations. As of the limitation of representation in the deep learning net-

work and widely occurring occlusion, it is impossible to interpret the 2D–3D relationship correctly using a single CNN model.

While the above-mentioned issues provide important clues for possible research directions, we believe that 3D object localization with hybrid representations [80,81] has considerable scope for improvement in the near future. Compared to unitary representation, a hybrid representation with edge, region, or creative geometric assumptions or any object-part awareness can use multiple training databases. Another possible direction is enforcing the consistency beyond diverse representations by training the network in a self-supervised manner. In particular, synthetic datasets can pave the way to robust representation for feature domain adaptation. Finally, beyond multiple intermediate representations, geometric relationships across object categories in different scenes can be ultimately formulated as an end-to-end optimization throughout an entire network. We believe that there is considerable scope for finding the best representation transforms, geometric relationships, or other physical conditions of 3D objects, and these discoveries can have a strong influence on future work.

6. Conclusions

Recently, deep learning methods have attracted considerable attention and witnessed rapid development. In contrast to previous hand-crafted features, the success of the CNN is attributed to its powerful ability to learn rich feature descriptions from an adequate amount of training data. Monocular 3D object detection is not an exception. Hence, we surveyed the current methodologies for deep-learning-based 3D object detection using single RGB images. They are being employed in various practical applications such as autonomous vehicles and robotics. We believe that the current gap between mature 2D-based methods and nascent 3D-based methods can be rapidly bridged on the basis of the intensive review presented herein. First, we summarized the widely used benchmark databases for training and evaluating the proposed methods in this area, and we reviewed the recent progress in monocular 3D object detection approaches by categorizing them into multi-stage and end-to-end approaches. We dealt with the main approaches used by recent methods to tackle the objective problem and discussed their underlying limitations. Finally, we examined the issues involved in localizing objects in the 3D space, which presently is an active research field because of its practical implications. Based on the current research status, object localization followed by pose estimation could be developed adequately for the 3D domain. In particular, enabling 3D perception only from a single camera will be useful for prospective applications.

Author Contributions: Conceptualization, S.-h.K. and Y.H.; formal analysis, S.-h.K. and Y.H.; investigation, S.-h.K. and Y.H.; data curation, S.-h.K.; writing—original draft preparation, S.-h.K.; writing—review and editing, S.-h.K. and Y.H.; supervision, Y.H.; funding acquisition, Y.H. Both authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2020-0-01462) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (number 2020R1F1A1077110).

Acknowledgments: The first author sincerely appreciates Min-ho Lee and Hoo-kyeong Lee at KETI for valuable discussion.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, W.; Luo, Y.; Wang, P.; Qin, Z.; Zhou, H.; Qiao, H. Recent Advances on Application of Deep Learning for Recovering Object Pose. In Proceedings of the 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), Qingdao, China, 3–7 December 2016; pp. 1273–1280.
2. Sahin, C.; Kim, T.K. Recovering 6D Object Pose: A Review and Multi-modal Analysis. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018; pp. 1–17.

3. Griffiths, D.; Boehm, J. A Review on Deep Learning Techniques for 3D Sensed Data Classification. *Remote Sens.* **2019**, *11*, 1499. [CrossRef]
4. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A Survey on 3D object Detection Methods for Autonomous Driving Applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
5. Wu, J.; Yin, D.; Chen, J.; Wu, Y.; Si, H.; Lin, K. A Survey on Monocular 3D Object Detection Algorithms Based on Deep Learning. *J. Phys. Conf. Ser.* **2020**, *1518*, 12–49. [CrossRef]
6. Rahman, M.M.; Tan, Y.; Xue, J.; Lu, K. Recent Advances in 3D Object Detection in the Era of Deep Neural Networks: A Survey. *IEEE Trans. Image Process.* **2019**, *29*, 2947–2962. [CrossRef] [PubMed]
7. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
10. Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; Sun, J. Light-head R-CNN: In Defense of Two-stage Object Detector. *arXiv* **2017**, arXiv:1711.07264.
11. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
12. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
13. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
14. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
15. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156.
16. Li, B.; Ouyang, W.; Sheng, L.; Zeng, X.; Wang, X. GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1019–1028.
17. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Subcategory-aware Convolutional Neural Networks for Object Proposals and Detection. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 924–933.
18. Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; Chateau, T. Deep MANTA: A Coarse-to-fine Many-task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2040–2049.
19. Manhardt, F.; Kehl, W.; Gaidon, A. ROI-10D: Monocular Lifting of 2D Detection to 6D Pose and Metric Shape. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2069–2078.
20. He, T.; Soatto, S. Mono3D++: Monocular 3D Vehicle Detection with Two-scale 3D Hypotheses and Task Priors. In Proceedings of the AAAI, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8409–8416.
21. Xu, B.; Chen, Z. Multi-level Fusion based 3D Object Detection from Monocular Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2345–2353.
22. Qin, Z.; Wang, J.; Lu, Y. MonoGRNet: A Geometric Reasoning Network for Monocular 3D Object Localization. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8851–8858.
23. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Oject Detection for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 8445–8453.
24. Roddick, T.; Kendall, A.; Cipolla, R. Orthographic Feature Transform for Monocular 3D Object Detection. In Proceedings of the British Machine Vision Conference (BMVC), Cardiff, UK, 9–12 September 2019, pp. 1–13.
25. Do, T.T.; Cai, M.; Pham, T.; Reid, I. Deep-6DPose: Recovering 6D Object Pose from a Single RGB Image. *arXiv* **2018**, arXiv:1802.10367.
26. Brazil, G.; Liu, X. M3D-RPN: Monocular 3D Region Proposal Network for Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 9287–9296.
27. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In Proceedings of the Robotics: Science and Systems (RSS), Pittsburgh, PA, USA, 26–30 June 2018; pp. 1–10. [CrossRef]
28. Tekin, B.; Sinha, S.N.; Fua, P. Real-time Seamless Single Shot 6D Object Pose Prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 292–301.

29. Kehl, W.; Manhardt, F.; Tombari, F.; Ilic, S.; Navab, N. SSD-6D: Making Rgb-based 3D Detection and 6D Pose Estimation Great Again. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1521–1529.
30. Chen, B.; Parra, A.; Cao, J.; Li, N.; Chin, T.J. End-to-end Learnable Geometric Vision by Back-propagating PnP Optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8100–8109.
31. Xiang, Y.; Mottaghi, R.; Savarese, S. Beyond Pascal: A Benchmark for 3D Object Detection in the Wild. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV), Steamboat Springs, CO, USA, 24–26 March 2014; pp. 75–82.
32. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (voc) Challenge. *IJCV* **2010**, *88*, 303–338. [CrossRef]
33. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.F. ImageNet: A Large-scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
34. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 567–576.
35. Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor Segmentation and Support Inference from RGBD Images. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; pp. 746–760.
36. Janoch, A.; Karayev, S.; Jia, Y.; Barron, J.T.; Fritz, M.; Saenko, K.; Darrell, T. A Category-level 3D Object Dataset: Putting the Kinect to Work. In *Consumer Depth Cameras for Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 141–165.
37. Xiao, J.; Owens, A.; Torralba, A. Sun3D: A Database of Big Spaces Reconstructed using SfM and Object Labels. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Darling Harbour, Sydney, Australia, 1–8 December 2013; pp. 1625–1632.
38. Xiang, Y.; Kim, W.; Chen, W.; Ji, J.; Choy, C.; Su, H.; Mottaghi, R.; Guibas, L.; Savarese, S. ObjectNet3D: A Large scale Database for 3D Object Recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 160–176.
39. Chang, A.X.; Funkhouser, T.; Guibas, L.; Hanrahan, P.; Huang, Q.; Li, Z.; Savarese, S.; Savva, M.; Song, S.; Su, H.; et al. ShapeNet: An Information-rich 3D Model Repository. *arXiv* **2015**, arXiv:1512.03012.
40. Tremblay, J.; To, T.; Birchfield, S. Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2038–2041.
41. Calli, B.; Walsman, A.; Singh, A.; Srinivasa, S.; Abbeel, P.; Dollar, A.M. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. *arXiv* **2015**, arXiv:1502.03143.
42. Hodan, T.; Michel, F.; Brachmann, E.; Kehl, W.; GlentBuch, A.; Kraft, D.; Drost, B.; Vidal, J.; Ihrke, S.; Zabulis, X.; et al. Bop: Benchmark for 6D Object Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–13 September 2018; pp. 19–34.
43. Hodaň, T.; Sundermeyer, M.; Drost, B.; Labbé, Y.; Brachmann, E.; Michel, F.; Rother, C.; Matas, J. BOP Challenge 2020 on 6D Object Localization. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 577–594.
44. Hinterstoisser, S.; Lepetit, V.; Ilic, S.; Holzer, S.; Bradski, G.; Konolige, K.; Navab, N. Model based Training, Detection and Pose Estimation of Texture-less 3D Objects in Heavily Cluttered Scenes. In Proceedings of the Asian Conference on Computer Vision (ACCV), Daejeon, Korea, 5–9 November 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 548–562.
45. Brachmann, E.; Krull, A.; Michel, F.; Gumhold, S.; Shotton, J.; Rother, C. Learning 6D Object Pose Estimation using 3D Object Coordinates. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 536–551.
46. Hodan, T.; Haluza, P.; Obdržálek, Š.; Matas, J.; Lourakis, M.; Zabulis, X. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. In Proceedings of the Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 880–888.
47. Drost, B.; Ulrich, M.; Bergmann, P.; Hartinger, P.; Steger, C. Introducing MVTec ITODD—A Dataset for 3D Object Recognition in Industry. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 2200–2208.
48. Kaskman, R.; Zakharov, S.; Shugurov, I.; Ilic, S. HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019; pp. 1–10.
49. Rennie, C.; Shome, R.; Bekris, K.E.; De Souza, A.F. A Dataset for Improved RGBD-based Object Detection and Pose Estimation for Warehouse Pick-and-place. *IEEE Robot. Autom. Lett.* **2016**, *1*, 1179–1185. [CrossRef]
50. Doumanoglou, A.; Kouskouridas, R.; Malassiotis, S.; Kim, T.K. Recovering 6D Object Pose and Predicting Next-best-view in the Crowd. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3583–3592.
51. Tejani, A.; Tang, D.; Kouskouridas, R.; Kim, T.K. Latent-class Hough Forests for 3D Object Detection and Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 462–477.

52. Ahmadyan, A.; Zhang, L.; Wei, J.; Ablavatski, A.; Grundmann, M. Objectron: A Large Scale Dataset of Object-Centric Videos in the Wild with Pose Annotations. *arXiv* **2020**, arXiv:2012.09988.
53. Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. The Open Images Dataset V4. *Int. J. Comput. Vis.* **2020**, *128*, 1956–1981 [CrossRef]
54. Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; Guibas, L.J. Normalized Object Coordinate Space for Category-level 6D Object Pose and Size Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2642–2651.
55. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
56. Cabon, Y.; Murray, N.; Humenberger, M. Virtual KITTI 2. *arXiv* **2020**, arXiv:2001.10773.
57. Gählert, N.; Jourdan, N.; Cordts, M.; Franke, U.; Denzler, J. Cityscapes 3D: Dataset and Benchmark for 9 DoF Vehicle Detection. *arXiv* **2020**, arXiv:2006.07864.
58. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
59. Wrenninge, M.; Unger, J. Synscapes: A Photo-realistic Synthetic Dataset for Street Scene Parsing. *arXiv* **2018**, arXiv:1810.08705.
60. Bengar, J.Z.; Gonzalez-Garcia, A.; Villalonga, G.; Raducanu, B.; Aghdam, H.H.; Mozerov, M.; Lopez, A.M.; van de Weijer, J. Temporal Coherence for Active Learning in Videos. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea, 27–28 October 2019; pp. 914–923.
61. Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; Lopez, A.M. The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3234–3243.
62. Su, H.; Qi, C.R.; Li, Y.; Guibas, L.J. Render for CNN: Viewpoint Estimation in Images using CNNs Trained with Rendered 3D Model Views. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 2686–2694.
63. Wohlhart, P.; Lepetit, V. Learning Descriptors for Object Recognition and 3D Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3109–3118.
64. Agarwal, S.; Mierle, K.;Bjorck, A.; Brown, D.C.; Byrd, R.H.; Chen, Y.; Conn, A.R.; Dellaer, F.; Golub, G.H.; Gould, N.; et al. Ceres Solver. Available online: http://ceres-solver.org (accessed on 20 January 2021).
65. Konishi, Y.; Hanzawa, Y.; Kawade, M.; Hashimoto, M. Fast 6D Pose Estimation from a Monocular Image using Hierarchical Pose Trees. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 398–413.
66. Muñoz, E.; Konishi, Y.; Murino, V.; Del Bue, A. Fast 6D Pose Estimation for Texture-less Objects from a Single RGB Image. In Proceedings of the International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 5623–5630.
67. Tjaden, H.; Schwanecke, U.; Schomer, E. Real-time Monocular Pose Estimation of 3D Objects using Temporally Consistent Local Color Histograms. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 124–132.
68. Mousavian, A.; Anguelov, D.; Flynn, J.; Kosecka, J. 3D Bounding Box Estimation using Deep Learning and Geometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7074–7082.
69. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep Ordinal Regression Network for Monocular Depth Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2002–2011.
70. Kim, Y.; Kum, D. Deep Learning based Vehicle Position and Orientation Estimation via Inverse Perspective Mapping Image. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 317–323.
71. Lepetit, V.; Moreno-Noguer, F.; Fua, P. EPnP: An Accurate O(N) Solution to the PnP Problem. *Int. J. Comput. Vis.* **2009**, *81*, 1–12. [CrossRef]
72. Peng, S.; Liu, Y.; Huang, Q.; Zhou, X.; Bao, H. PVNet: Pixel-wise Voting Network for 6DOF Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4561–4570.
73. Poirson, P.; Ammirato, P.; Fu, C.Y.; Liu, W.; Kosecka, J.; Berg, A.C. Fast Single Shot Detection and Pose Estimation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 676–684.
74. Liu, L.; Lu, J.; Xu, C.; Tian, Q.; Zhou, J. Deep Fitting Degree Scoring Network for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 1057–1066.
75. Li, Y.; Wang, G.; Ji, X.; Xiang, Y.; Fox, D. DeepIM: Deep Iterative Matching for 6D Pose Estimation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 683–698.

76. Rad, M.; Lepetit, V. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without using Depth. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3828–3836.
77. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
78. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-driven 6D Object Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3385–3394.
79. Zakharov, S.; Shugurov, I.; Ilic, S. DPOD: 6D Pose Object Detector and Refiner. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–3 November 2019; pp. 1941–1950.
80. Hodan, T.; Barath, D.; Matas, J. EPOS: Estimating 6D Pose of Objects With Symmetries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 11703–11712.
81. Song, C.; Song, J.; Huang, Q. HybridPose: 6D Object Pose Estimation under Hybrid Representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 431–440.