applied sciences

MDPI

Article

Object Pose Detection to Enable 3D Interaction from 2D Equirectangular Images in Mixed Reality Educational Settings

Matteo Zanetti [1,*], Alessandro Luchetti [1], Sharad Maheshwari [1], Denis Kalkofen [2], Manuel Labrador Ortega [3] and Mariolino De Cecco [1]

[1] Department of Industrial Engineering, University of Trento, Via Sommarive 9, 38123 Trento, Italy; alessandro.luchetti@unitn.it (A.L.); sharad.maheshwari@studenti.unitn.it (S.M.); mariolino.dececco@unitn.it (M.D.C.)
[2] Institute of Computer Graphics and Vision, Graz University of Technology, Rechbauerstraße 12, 8010 Graz, Austria; kalkofen@icg.tugraz.at
[3] Resources Innovation Center Leoben, Montanuniversität Leoben, Franz Josef Strasse 18, 8700 Leoben, Austria; manuel.labrador-ortega@unileoben.ac.at
* Correspondence: matteo.zanetti@unitn.it; Tel.: +39-0461-285263

Abstract: In this paper, we address the challenge of estimating the 6DoF pose of objects in 2D equirectangular images. This solution allows the transition to the objects' 3D model from their current pose. In particular, it finds application in the educational use of 360° videos, where it enhances the learning experience of students by making it more engaging and immersive due to the possible interaction with 3D virtual models. We developed a general approach usable for any object and shape. The only requirement is to have an accurate CAD model, even without textures of the item, whose pose must be estimated. The developed pipeline has two main steps: vehicle segmentation from the image background and estimation of the vehicle pose. To accomplish the first task, we used deep learning methods, while for the second, we developed a 360° camera simulator in Unity to generate synthetic equirectangular images used for comparison. We conducted our tests using a miniature truck model whose CAD was at our disposal. The developed algorithm was tested using a metrological analysis applied to real data. The results showed a mean difference of 1.5° with a standard deviation of 1° from the ground truth data for rotations, and 1.4 cm with a standard deviation of 1.5 cm for translations over a research range of ±20° and ±20 cm, respectively.

Keywords: image processing; 6DoF pose estimation; mixed reality; human empowerment; educational setting

Citation: Zanetti, M.; Luchetti, A.; Maheshwari, S.; Kalkofen, D.; Ortega, M.L.; De Cecco, M. Object Pose Detection to Enable 3D Interaction from 2D Equirectangular Images in Mixed Reality Educational Settings. *Appl. Sci.* **2022**, *12*, 5309. https://doi.org/10.3390/app12115309

Academic Editor: Jorge Martin-Gutierrez

Received: 20 April 2022
Accepted: 23 May 2022
Published: 24 May 2022

1. Introduction

Mixed Reality (MR) has the potential to increase the learning experiences in several scenarios [1] such as industrial maintenance training [2,3]; in Industry 4.0, cyber-physical engineering can be made experienceable and tangible [4]; construction training stressing safety aspects using Augmented Virtuality to enhance hazard recognition [5,6]; and humanities disciplines such as history [7], medicine [8], public speaking [9], and others. In general, MR technologies allow students to "visit" places, such as production plants, power plants, mine sites, and plants for which special permits are mandatory. Moreover, there are often limited resources for practical lessons in the education sector. MR technologies can help to overcome these limits and improve learning. In Europe, there are already education projects that propose this new paradigm, such as MiReBooks [10]. MiReBooks aims to produce a series of Virtual Reality (VR) and Augmented Reality (AR) based interactive mining handbooks as a new digital standard for higher mining education across Europe. Current challenges in mining education are met in an innovative new way, combining classical paper-based teaching materials with MR materials and their transformation into pedagogically and didactically coherent MR handbooks for integrative classroom use [11].

From a technological point of view, it is possible to use immersive solutions that exploit VR gears that are able to display virtual environments augmented with real content, a real environment fully reconstructed in 3D or, enabled by a modern high-resolution whole field of view cameras—360° videos. A great effort is needed to accurately replicate real 3D environments in virtual scenarios: acquiring 360° videos is a much simpler process. On the other hand, 360° videos are not able to provide 3D perception.

For this reason, we propose a method to generate 2D to 3D transitions starting from 360° source equirectangular images: when an interaction with an object is needed, we can propose to the user the 3D virtual model of the latter. To accomplish this, there is the need to develop a robust algorithm capable of estimating the object's pose inside the 360° frame. 6DoF estimation is one of the main challenging research topics in computer vision [12–14].

Specifically, in this study we present a novel pipeline whose aim is to infer the pose of a vehicle given a single 2D equirectangular image. The proposed algorithms are described in Section 5, and the experimental results are provided in Section 6. Although the exploitation is very general, we furthermore demonstrate the application of our approach. Therefore, in Section 2 we show the effects of human neuropsychological empowerment through MR, and we provide an example implementation in Section 3, which demonstrates the proposed pipeline for enabling 2D to 3D transitions in MR environments.

2. Human Neuropsychological Empowerment in Learning Scenarios through Mixed Reality

While MR can enhance the human capabilities in several contexts such as clinics [15], personalized support for mild cognitive impaired users in cooking [16], cognitive training system for mild cognitive impairment [17], can enhance the human-robot interaction while sharing the same virtual-real (mirrored) environment [18]. Regarding the interaction channel, MR can fully exploit a natural interface [19,20] while as input MR can exploit a wide set of channels such as visual text, 2D, 3D, auditory, and kinesthetics [21]. In training, a recent meta-analysis [22] shows that whether one trains in a virtual or a real setting, the results are essentially equivalent. In teaching, MR can enhance the learner's neuropsychological functions during the comprehension and memorization of the educational concepts. More specifically, in the following, we take into consideration the human perception-action loop (interface and interaction modality) and the whole patterns of neuropsychological functions highlighting the impact of each of them.

Regarding Orientation, immersive MR simulations allow the understanding of complex 3D phenomena that would be difficult to comprehend through other media [23,24]. Collaborative MR can transmit procedural knowledge in manufacturing training almost perfectly, replacing other forms of face-to-face training [25].

Regarding Attention, within MR, information can be spatially and temporally aligned with physical items and the learners' activities. This solves the problem of the human brain, which has a limited capacity for processing information from sensory channels (too much information results in cognitive overload/poor performances of the selective function and thus poor quality in learning). In fact, in MR it is possible to enhance the spatial and the temporal contiguity effect that makes students learn better when multiple representations of the same information are presented closer in space rather than far apart and/or when multiple representations of the same information are presented at the same time, rather than separated in time [26]. This contextualizes the different information content in an aggregated form within the learning scenario and uses different sensor modalities such as auditory, visual, and tactile, which are processed in parallel within our brain. Furthermore, the "digital augmentation" of reality can direct user attention toward the relevant content. This mechanism effectively enhances learning tasks involving visuospatial information. The system presented in [27] highlights organs to effectively teach students about human anatomy.

Regarding Memory, an enhanced memory encoding is stimulated by the physical immersion of MR experiences, and the fact that users interact with their senses, body,

and limbs, which cause learners to encode sensing and proprioceptive information along with the educational content [28].

Regarding Language, MR can leverage current translation technologies to temporally align the verbal translated content to the user.

Regarding Visual perception, recent advances in display technologies are attempting to fill the gap between the natural view and the one mediated by the display. This is not the appropriate place to focus on this topic.

Regarding Motion planning and control, MR can increase depth perception as in [29] where the perception of 3D obstacles is enhanced via 2D projection onto the ground plane with perspective correction based on the subject's head position. Another way MR can be of benefit is augmenting user exproprioception. Visual exproprioception provides information about the body position in relation to the environment, and it can yield positive effects on position control and gait biomechanics in AR systems. As an example, in [30] an AR application was implemented using Microsoft HoloLens. The experiment revealed that the interface projected in front of the user (instead of the ground), and, from a third-person point of view, improves posture, visual stimuli, and safety.

3. 2D to 3D Objects Transitioning and Its Use in MR Settings

The 360° cameras support capturing dynamic environments from a single point of view. In addition, VR headsets enable viewing the captured environments by supporting head rotations around all three axes. While this enables immersive experiences, the missing translations may cause several perceptual issues [31], and it limits explorations to the pre-defined viewpoint. To enable full 3D object explorations, we propose combining 2D panoramic video data with renderings of registered 3D models. In addition, we propose providing a transitional interface [32,33] for switching between the 2D and 3D data. Thus, an essential requirement for a smooth transition is that the object must first be identified, localized in 6DoF, and augmented with its 3D model when it transitions from video to 3D. Then, after transitioning to the 3D virtual environment, the model can be freely explored (Figure 1).

(**a**) (**b**) (**c**) (**d**)

Figure 1. Transitioning from 2D video to the 3D virtual object: (**a**) 2D video. (**b**) Object replacement after detection and localization. (**c**) Object rotating in front of the user viewpoint. (**d**) Digital information contextualized with the vehicle model. The corresponding videos can be found here (https://youtu.be/E1fwuexrCo0, accessed on 20 April 2022).

4. Related Work

6DoF pose estimation using RGB images involves different fields such as bin picking problems [34], robot manipulation [35], autonomous vehicles [36], and MR applications [37].

Usually, to accomplish this task, deep learning methods are used. One of the main approaches to 6DoF pose estimation, as described in [38], is to decouple the translation and the rotation estimation. The translation is estimated by localizing the centre of the object in the image and predicting its distance from the camera. After that, the rotation is estimated by regressing to a quaternion representation. A 6DoF Object detection system with two stages is also proposed in [39]. A single Shot Multibox Detector (SSD) [40] extracts the object bounding boxes and an Augmented AutoEncoder (AAE) estimates the object rotation. Similar to the previous approach, DCS-PoseNet [41] uses a two-step process to estimate 6DoF from 2D object bounding-boxes. First, the framework segments the object from the cropped image, followed by predicting 6DoF pose using DSC-PoseNet, which employs a differential renderer.

Some solutions try to regress rotation and translation simultaneously. For example, 6D-VNet [42] uses an end-to-end deep learning network to estimate the 6DoF pose of vehicles. The network extends the Mask R-CNN object detector and takes its intermediate outputs and further regresses for rotation and translation of the object in 3D space.

Other approaches instead try to solve a Perspective-n-Point problem [43]. For example, the pose estimation method Pix2Pose [44] proposes a deep learning network to supplement a 2D detection pipeline to enable pose estimation. It regresses pixel-wise 3D coordinates from images using texture-less 3D models. The pixel-wise prediction is used to form 2D–3D correspondences. Finally, the PnP algorithm can be applied.

In Ref. [45], the authors propose an extension of the EfficientDet architecture [46] used for 2D object detection to predict the rotation and the translation of the object in the 3D space.

Most current works describe the problem statement and solution for regular RGB images. The application of the algorithms of these works to equirectangular images is tricky. The main reason is that equirectangular images present severe distortion and there is a lack of training data related to these types of images. To the best of our knowledge, some works try to perform 2D object detection in equirectangular images [47,48], but none perform an estimate of the 6DoF pose.

For this reason, our work contributes moving toward objects pose estimation also for this kind of images. Indeed, 360° videos are becoming more popular in the educational context, as they provide students with a more engaging learning activity [49]. The method presented in this paper shows an example of how this media, together with VR technologies, can create contemporary virtual learning environments, which enable students to experience and interact with virtual content [50]. 360° videos, in contrast with 2D movies, offer an immersive experience not only by giving the perception of being physically present in a virtual environment but also allowing interaction with objects. This is only possible when knowing the pose of the object with which we want to interact with inside the scene.

Our pipeline specifically solves the problem of 6DoF pose estimation for objects in equirectangular images. Additionally, while other methods primarily rely on deep learning models to perform the task, ours uses deep learning only for segmentation, which is the first step. It then uses an optimization technique for pose estimation that does not require a trained network and that can be applied with no effort to any type of object. Indeed, a benefit of the proposed method, in comparison to the related works presented in this section, is that there is no need to create a training dataset for the pose estimation, avoiding a task that can be quite a time consuming and difficult in terms of the acquisition of the ground truth pose, scalability, and full coverage of possible poses [51]. Other approaches use synthetic data for the training to avoid the this cumbersome task [52,53]. However, this method has the handicap of the gap domain between real and synthetic data [54]. To resume, in comparison to the traditional methods, the object pose detection phase of our method can be directly tested on the data that must be processed, ensuring a considerable time saving.

5. Algorithm Description

We can subdivide the developed algorithm into two main steps:

- Vehicle segmentation from the equirectangular image;
- Vehicle pose detection with respect to the camera reference frame (6DoF).

To accomplish the first task, we made use of Yolact++ [55], a convolutional model for real-time instance segmentation. Yolact++ proved to be pretty accurate for the segmentation of trucks, also in equirectangular images, in which the distortion is significant. However, to make the vehicle segmentation more robust, we trained Yolact++ with 500 equirectangular images by manually labelling trucks in frames taken from 360° videos of open-pit mining operations. Figure 2 shows the result of the trained Yolact++ model.

(**a**) (**b**)

Figure 2. The result of the trained convolutional model Yolact++ for an equirectangular image of a truck. (**a**) An example of an input image showing a truck in a mining environment. (**b**) Result of the segmentation in which the truck is correctly segmented.

The only requirement for the vehicle pose detection is to have an accurate CAD model even without textures of the item, whose pose must be estimated. For the tests described in this paper, we used the CAD model of a Komatsu HD785 truck. We used the CAD model inside Unity 3D, a cross-platform game engine. In this case, we used Unity to create a simulation environment. We implemented a 360° camera simulator in this environment to capture equirectangular images of the CAD model.

Using the output given by Yolact++, we can perform the vehicle pose detection. To accomplish this task, we developed the algorithm schematically described in Figure 3.

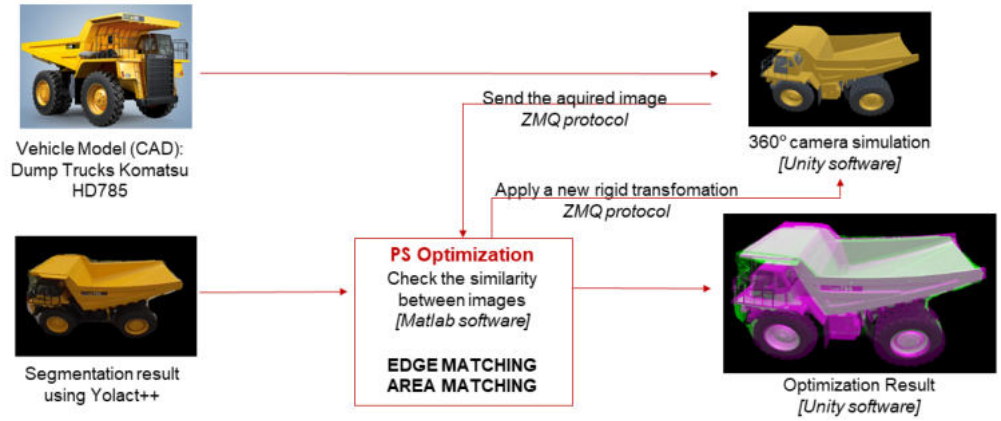


Figure 3. Scheme of the pose detection algorithm.

Through a connection based on the ZMQ protocol, we allow data exchange between Unity and Matlab. We put the vehicle in a random pose inside Unity. Then we take a picture of the scene using the 360° camera simulation. We will call this picture the synthetic image. In Matlab, we compare this synthetic image with the output of Yolact++ (the segmented

real-world image). A score is given for the similarity between the two pictures. A Particle Swarm (PS) optimizer is in charge of finding the optimal solution. At each new iteration, the algorithm sends a new pose to Unity. The 360° simulated camera takes a new picture, and we repeat the previous steps until convergence or the maximum number of iterations are reached.

Hereafter, the expression of the Cost Function (CF):

$$CF = S_E + S_A + S_C + S_V, \tag{1}$$

where its terms depend on:

- Edges (S_E);
- Area (S_A);
- Difference in the centroids of the edges (S_C);
- Difference in the eigen vectors of the edges (S_V);

In the following subsections, we explain the various term of the cost function in detail. It is important to highlight that the terms of the cost function also work for texture-less objects because they mainly rely on shape features than on texture. To show the potential of the developed algorithm, we choose to apply it to a vehicle model whose CAD was at our disposal. The CAD model does not perfectly match the real object in every detail but we will show that the results are accurate enough.

The real-world and synthetic images in Figure 4 are taken as an example to show the computations made for the different terms of the cost function.

(**a**) (**b**)

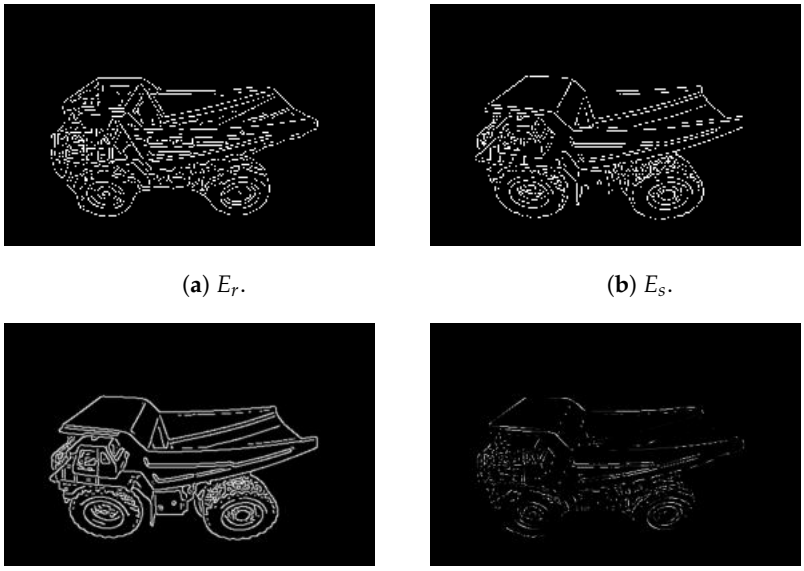
Figure 4. Real-world and synthetic images taken as example to illustrate the different terms of the cost function. (**a**) Real-world image. (**b**) Synthetic image.

5.1. Edges

The first term of the cost function is relative to edges. Using the "Canny" algorithm [56], we computed the images of the edges of the real-world (E_r) and synthetic image (E_s). Since a perfect correspondence between the CAD model and the actual vehicle is impossible, we smoothed the edges of E_s by applying a Gaussian filter with a standard deviation of 0.5. We will call this last image E_{sg}. The two images are then multiplied pixel by pixel, computing E_m as:

$$E_m = E_r \cdot E_{sg}. \tag{2}$$

Figure 5 shows the images involved in the computation.



(**a**) E_r. (**b**) E_s.

(**c**) E_{sg}. (**d**) E_m.

Figure 5. The images involved in the computation of the cost function term relative to edges. (**a**) E_r, edges of the real-world image. (**b**) E_s, edges of the synthetic image. (**c**) E_{sg}, edges of the synthetic image after the Gaussian filter. (**d**) E_m, pixel-wise multiplication between E_r and E_{sg}.

The score term is computed as follows:

$$S_E = 1.0 - \frac{n_m}{n_s}, \tag{3}$$

where n_m and n_s are the number of pixels of E_m and E_s, which are greater than zero.

5.2. Areas

The corresponding binary images (BW_r and BW_s) are computed from the real-world and the synthetic ones. We call A_s the area of BW_s, which is the number of pixels whose value is greater than 0. A dilated version of BW_s, called BW_{sd}, is also computed using a disk with a diameter of 7 pixels as the morphological structuring element. Let us indicate with BW_d the difference between BW_{sd} and BW_s:

$$BW_d = BW_{sd} - BW_s. \tag{4}$$

A_d is the area of BW_d.

Now, it is possible to compute the images M_a and M_d, i.e., the result of the pixel-wise multiplication between BW_r and BW_s, and between BW_r and BW_d:

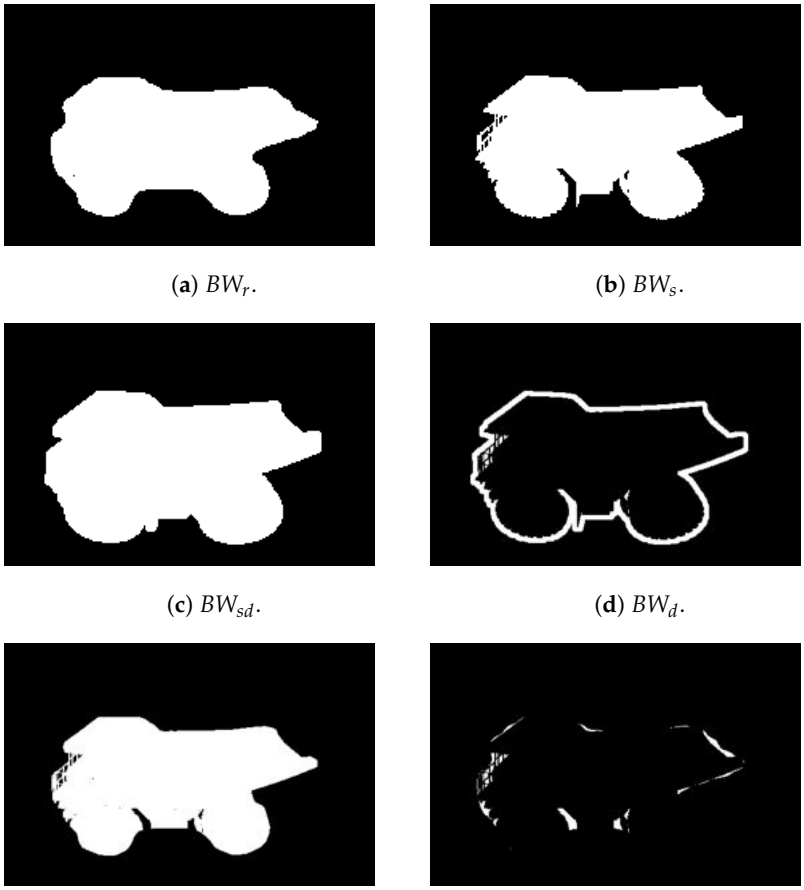
$$M_{rs} = BW_r \cdot BW_s, \tag{5}$$
$$M_{rd} = BW_r \cdot BW_d. \tag{6}$$

Figure 6 shows the images involved in the computations.

We can compute the score relative to the areas with the following equation:

$$S_A = 1.0 - A_{rs}/A_s + A_{rd}/A_d, \tag{7}$$

where A_{rs} and A_{rd} are the corresponding areas of M_{rs} and M_{rd}.



(**a**) BW_r. (**b**) BW_s.

(**c**) BW_{sd}. (**d**) BW_d.

(**e**) M_{rs}. (**f**) M_{rd}.

Figure 6. The images involved in the computation of the cost function term relative to the areas. (**a**) BW_r, binary image of the real-world image. (**b**) BW_s, binary image of the synthetic image. (**c**) BW_{sd}, synthetic binary image dilated. (**d**) BW_d, result of the subtraction between BW_{sd} and BW_s. (**e**) M_{rs}, result of the multiplication between BW_r and BW_s. (**f**) M_{rd}, result of the multiplication between BW_r and BW_d.

5.3. Difference in the Centroids of the Edges

This part of the cost function is in charge of computing the difference between the centroids of the images E_r and E_m. The formula to compute the centroids of the images is as follows:

$$x_c = \frac{\sum_{i=1}^{N} I(x_i, y_i) \cdot x_i}{\sum_{i=1}^{N} I(x_i, y_i)}, \tag{8}$$

$$y_c = \frac{\sum_{i=1}^{N} I(x_i, y_i) \cdot y_i}{\sum_{i=1}^{N} I(x_i, y_i)}, \tag{9}$$

where (x_c, y_c) are the coordinates of the centroid of the image I, N is the number of pixels whose value is greater than 0, (x_i, y_i) are the general coordinates of the pixel i, and $I(x_i, y_i)$ is the grey value of the pixel in position (x_i, y_i).

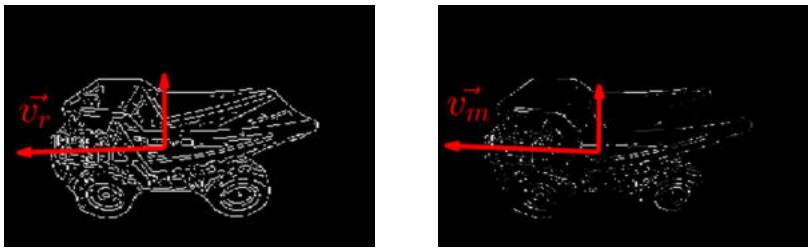
The cost function term is computed as:

$$S_C = \frac{\sqrt{(x_{cr} - x_{cm})^2 + (y_{cr} - y_{cm})^2}}{\sqrt{R^2 + C^2}}, \tag{10}$$

where (x_{cr}, y_{cr}) and (x_{cm}, y_{cm}) are the coordinates of the centroids of E_r and E_m, and R and C are the number of rows and columns of E_r.

5.4. Difference in the Eigen Vectors of the Edges

The last term of the cost function can be explained as a constraint for the edge matching to be uniform on all the parts of the edge images, i.e., E_r and E_{sg}. To reach this aim, we can arrange the image coordinates (x_i, y_i) of the pixels whose value is greater than 0 in a matrix of dimension $N \times 2$, where N is the number of pixels whose value is greater than 0. Then we can compute the covariance matrices C_r and C_m of this matrix for E_r and E_m. Once C_r and C_m are obtained, we can compute the eigenvectors for both. Let us call $\vec{v_r}$ and $\vec{v_m}$ the two eigenvectors that correspond to the highest eigenvalue for C_r and C_m (Figure 7).



(**a**) Eigenvector representation of E_r. (**b**) Eigenvector representation of E_m.

Figure 7. E_m and E_r with their respective eigenvectors centred in the centroids of the two images. (**a**) E_r and its eigenvectors. (**b**) E_m and its eigenvectors.

The cost function term is the dot product between $\vec{v_r}$ and $\vec{v_m}$:

$$S_V = 1.0 - \vec{v_r} \cdot \vec{v_m}. \tag{11}$$

6. Results

Figure 8 shows the experimental setup used to test the developed algorithm. A 360° camera, such as the Insta360 ONE X, is placed on a rotary stage which in turn is placed on a translation stage. The camera frames a miniature model of a Komatsu HD785 truck.

Figure 8. Experimental setup to test the developed algorithm. A 360° camera is placed on a rotary and a translation stage. The camera frames the miniature model of a truck.

We acquired 10 images by translating the translation stage 8 cm of each new acquisition, and 11 images by rotating the camera 5° of each new acquisition (Figure 9).

Concerning the parameters used for the PS optimization, we set the swarm size to 150 and the maximum number of iterations to 75. The research range was set to ±20° for rotations and to ±20 cm for translations. The initial pose conditions were set randomly from nominal values within the imposed research ranges. The algorithm ran on an Intel(R)

Core(TM) i7-9700KF CPU. The mean computational time to find the optimum was about 20 min.

Table 1 and Figure 10 show the results obtained for the imposed rotations.

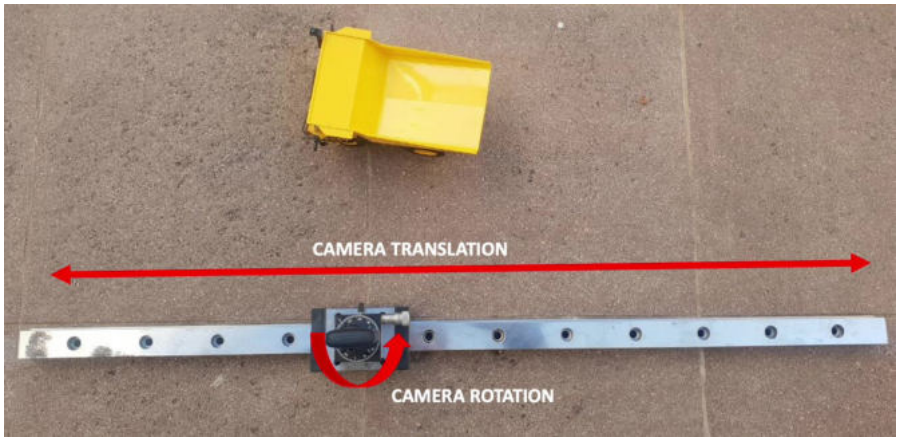


Figure 9. Scheme of the rotations and translations imposed to the camera.

Table 1. Results obtained by the algorithm, applying a rotation of 5° at each step.

Nominal Angle [°]	Measured Angle [°]	Difference [°]
5.0	6.4	1.4
10.0	10.1	0.1
15.0	17.3	2.3
20.0	19.7	−0.3
25.0	26.3	1.3
30.0	31.0	1.0
35.0	37.0	2.0
40.0	41.9	1.9
45.0	48.2	3.2
50.0	52.1	2.1

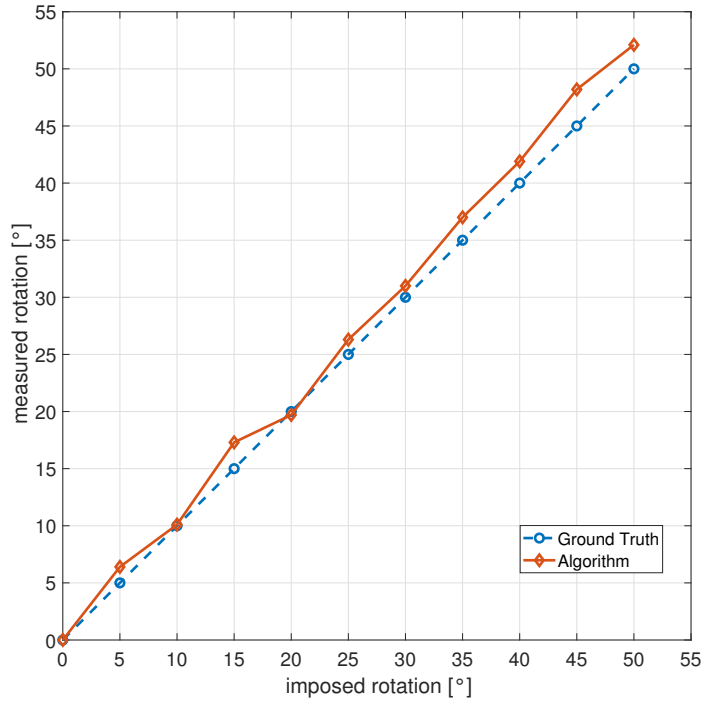


Figure 10. Comparison of the imposed rotations with the measured ones.

Table 2 and Figure 11 show the results obtained for the imposed translations.

Table 2. Results obtained by the algorithm applying a translation of 8 cm at each step.

Nominal Translation [cm]	Measured Translation [cm]	Difference [cm]
8.0	8.4	0.4
16.0	16.0	0.0
24.0	24.9	0.9
32.0	32.9	0.9
40.0	40.8	0.8
48.0	49.0	1.0
56.0	57.1	1.1
64.0	67.9	3.9
72.0	76.0	4.0

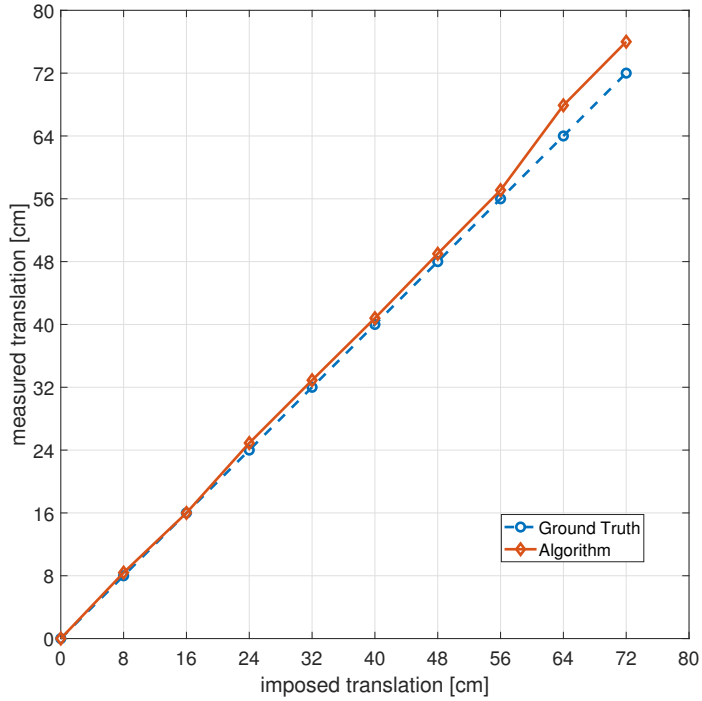
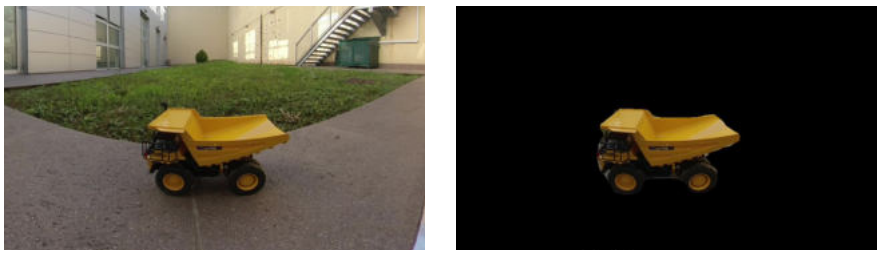


Figure 11. Comparison of the imposed translations with the measured ones.

Figure 12 shows an example of the results obtained; in this case the camera was rotated by 15° with respect to the initial orientation.



(**a**) (**b**)

Figure 12. *Cont.*

(**c**)

Figure 12. An example of the optimization result where the camera was rotated by 15°. (**a**) Portion of the input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result in which the CAD is rendered in the final pose found by the optimization algorithm.

7. Discussion

Results show that the developed algorithm achieved good results for both translations and rotations. In particular, the maximum difference in the rotation estimation was 3.2° for the nominal rotation of 45° (see Table 1). Figure 13 shows the optimization result for this case.

(**a**)

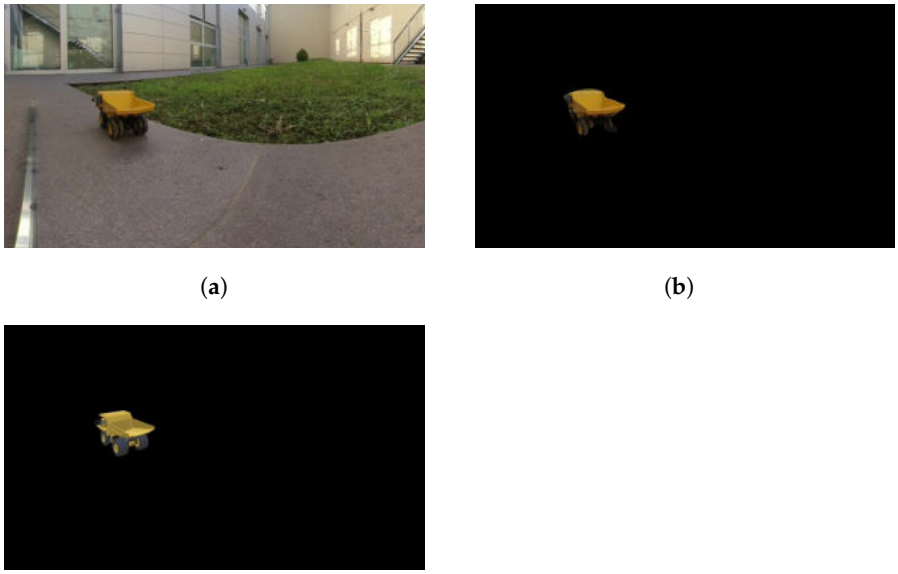
(**b**)

(**c**)

Figure 13. Optimization result where the camera was rotated by 45°. (**a**) Input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result in which the CAD is rendered in the final pose found by the optimization algorithm.

The mean difference for the rotation is 1.5°, while the standard deviation is 1.0°.

The maximum difference in the translation estimation was instead 4 cm for the nominal translation of 72 cm (see Table 2). As shown in Figure 11, it seems that, at the increase of the amount of the translation, the difference increases. Most likely, looking at Figure 14, this is due to how the vehicle appears in the equirectangular image. In this case, the vehicle appears quite far and small translations cannot be appreciated from the image point of view. Indeed, also in this case, at least visually, the difference between the real world and the synthetic image does not seem to be relevant (Figure 14).



(**a**) (**b**)

(**c**)

Figure 14. Optimization result where the camera was translated by 72 cm. (**a**) Input equirectangular image taken by the 360° camera. (**b**) Result of the segmentation. (**c**) Optimization result in which the CAD is rendered in the final pose found by the optimization algorithm.

The mean difference for the translation is 1.4 cm, while the standard deviation is 1.5 cm. The computational time of 20 min makes use of the proposed algorithm to be used offline. However, in the case of a video, once the pose is estimated in the first frame, the search field for the next frame is very limited because the vehicle will be in a pose very near to one of the previous frames. This will speed up the elaboration and the pose detection.

8. Conclusions

We presented an innovative method to estimate the 6DoF pose of vehicles in equirectangular images. This method relies on deep learning methods only for the object segmentation, while the pose is estimated through a cost function optimization. Only the CAD model of the object is needed for this step even without textures for the nature of the cost function used. This makes our method quite flexible to be applied to any kind of object and lighting conditions due to the lack of colour-affected terms in the comparison for pose estimation. We tested the results of our algorithm through an experimental setup, comparing them to measured rotations and translations applied to the camera in the real world. We obtained a maximum difference of 3.2° from the ground truth data for rotations, and 4 cm for translations over a research range of $\pm 20°$ and ± 20 cm, respectively. Future works will try to improve the computational time and reduce the pose detection error.

Author Contributions: Conceptualization, M.Z., A.L., D.K., M.L.O. and M.D.C.; formal analysis, M.Z. and A.L.; investigation, M.Z., A.L. and S.M.; writing—original draft preparation, M.Z., A.L., D.K. and M.D.C.; supervision, M.D.C., M.L.O. and D.K.; project administration, M.D.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Institute of Innovation and Technology (EIT) Raw Materials in the project MiReBooks: Mixed Reality Handbooks for Mining Education (18060).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Dunleavy, M.; Dede, C. Augmented reality teaching and learning. In *Handbook of Research on Educational Communications and Technology*; Springer: Berlin/Heidelberg, Germany, 2014.
2. Borsci, S.; Lawson, G.; Broome, S. Empirical evidence, evaluation criteria and challenges for the effectiveness of virtual and mixed reality tools for training operators of car service maintenance. *Comput. Ind.* **2015**, *67*, 17–26. [CrossRef]
3. Neges, M.; Adwernat, S.; Abramovici, M. Augmented virtuality for maintenance training simulation under various stress conditions. *Procedia Manuf.* **2018**, *19*, 171–178. [CrossRef]
4. Quint, F.; Sebastian, K.; Gorecky, D. A mixed-reality learning environment. *Procedia Comput. Sci.* **2015**, *75*, 43–48. [CrossRef]
5. Chen, A.; Golparvar-Fard, M.; Kleiner, B. Design and development of SAVES: A construction safety training augmented virtuality environment for hazard recognition and severity identification. In *Computing in Civil Engineering (2013)*; ASCE: Reston, VA, USA, 2013; pp. 841–848.
6. Albert, A.; Hallowell, M.R.; Kleiner, B.; Chen, A.; Golparvar-Fard, M. Enhancing construction hazard recognition with high-fidelity augmented virtuality. *J. Constr. Eng. Manag.* **2014**, *140*, 04014024. [CrossRef]
7. Gheorghiu, D.; Stefan, L. Augmented Virtuality as an instrument for a better learning of history. In Proceedings of the 13th International Conference on Virtual Learning (ICVL 2018), Alba Iulia, Romania, 26–27 October 2018; pp. 26–27.
8. Jamali, S.S.; Shiratuddin, M.F.; Wong, K.W.; Oskam, C.L. Utilising mobile-augmented reality for learning human anatomy. *Procedia-Soc. Behav. Sci.* **2015**, *197*, 659–668. [CrossRef]
9. Zhou, H.; Fujimoto, Y.; Kanbara, M.; Kato, H. Virtual Reality as a Reflection Technique for Public Speaking Training. *Appl. Sci.* **2021**, *11*, 3988. [CrossRef]
10. Daling, L.; Kommetter, C.; Abdelrazeq, A.; Ebner, M.; Ebner, M. Mixed Reality Books: Applying Augmented and Virtual Reality in Mining Engineering Education. In *Augmented Reality in Education*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 185–195.
11. Kalkofen, D.; Mori, S.; Ladinig, T.; Daling, L.; Abdelrazeq, A.; Ebner, M.; Ortega, M.; Feiel, S.; Gabl, S.; Shepel, T.; et al. Tools for Teaching Mining Students in Virtual Reality based on 360°Video Experiences. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 22–26 March 2020; pp. 455–459. doi: 10.1109/VRW50115.2020.00096. [CrossRef]
12. He, Z.; Feng, W.; Zhao, X.; Lv, Y. 6D Pose Estimation of Objects: Recent Technologies and Challenges. *Appl. Sci.* **2021**, *11*, 228. [CrossRef]
13. Tremblay, J.; To, T.; Sundaralingam, B.; Xiang, Y.; Fox, D.; Birchfield, S. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv* **2018**, arXiv:1809.10790.
14. Hoque, S.; Arafat, M.Y.; Xu, S.; Maiti, A.; Wei, Y. A Comprehensive Review on 3D Object Detection and 6D Pose Estimation With Deep Learning. *IEEE Access* **2021**, *9*, 143746–143770. [CrossRef]
15. De Cecco, M.; Fornaser, A.; Tomasin, P.; Zanetti, M.; Guandalini, G.; Ianes, P.; Pilla, F.; Nollo, G.; Valente, M.; Pisoni, T. Augmented reality to enhance the clinician's observation during assessment of daily living activities. In Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics, 12–15 June, 2017; Springer: Cham, Switzerland, 2017; pp. 3–21.
16. D'Agostini, J.; Bonetti, L.; Salem, A.; Passerini, L.; Fiacco, G.; Lavanda, P.; Motti, E.; Stocco, M.; Gashay, K.; Abebe, E.; et al. An augmented reality virtual assistant to help mild cognitive impaired users in cooking a system able to recognize the user status and personalize the support. In Proceedings of the IEEE 2018 Workshop on Metrology for Industry 4.0 and IoT, Brescia, Italy, 16–18 April 2018; pp. 12–17.
17. Park, E.; Yun, B.J.; Min, Y.S.; Lee, Y.S.; Moon, S.J.; Huh, J.W.; Cha, H.; Chang, Y.; Jung, T.D. Effects of a mixed reality-based cognitive training system compared to a conventional computer-assisted cognitive training system on mild cognitive impairment: A pilot study. *Cogn. Behav. Neurol.* **2019**, *32*, 172–178. [CrossRef]
18. Hoenig, W.; Milanes, C.; Scaria, L.; Phan, T.; Bolas, M.; Ayanian, N. Mixed reality for robotics. In Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 28 September–2 October 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5382–5387.
19. Conci, N.; Ceresato, P.; De Natale, F.G. Natural human-machine interface using an interactive virtual blackboard. In Proceedings of the 2007 IEEE International Conference on Image Processing, San Antonio, TX, USA, 16 September–19 October 2007; IEEE: Piscataway, NJ, USA, 2007; Volume 5, pp. V–181.
20. Aliprantis, J.; Konstantakis, M.; Nikopoulou, R.; Mylonas, P.; Caridakis, G. Natural Interaction in Augmented Reality Context. In Proceedings of the VIPERC@ IRCDL, Pisa, Italy, 30 January 2019; pp. 50–61.
21. Radu, I. Augmented reality in education: A meta-review and cross-media analysis. *Pers. Ubiquitous Comput.* **2014**, *18*, 1533–1543. [CrossRef]
22. Kaplan, A.D.; Cruit, J.; Endsley, M.; Beers, S.M.; Sawyer, B.D.; Hancock, P. The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Hum. Factors* **2021**, *63*, 706–726. [CrossRef]
23. Kaufmann, H.; Meyer, B. Simulating educational physical experiments in augmented reality. In Proceedings of the ACM SIGGRAPH Asia 2008 Educators Programme, Singapore, 10–13 December 2008; ACM Publications: New York, NY, USA, 2008; pp. 1–8.

24. Baldassi, S.; Cheng, G.T.; Chan, J.; Tian, M.; Christie, T.; Short, M.T. Exploring immersive AR instructions for procedural tasks: The role of depth, motion, and volumetric representations. In Proceedings of the 2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct), Merida, Mexico, 19–23 September 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 300–305.
25. Gonzalez-Franco, M.; Pizarro, R.; Cermeron, J.; Li, K.; Thorn, J.; Hutabarat, W.; Tiwari, A.; Bermell-Garcia, P. Immersive mixed reality for manufacturing training. *Front. Robot. AI* **2017**, *4*, 3. [CrossRef]
26. Mayer, R.E.; Moreno, R. Nine ways to reduce cognitive load in multimedia learning. *Educ. Psychol.* **2003**, *38*, 43–52. [CrossRef]
27. Nischelwitzer, A.; Lenz, F.J.; Searle, G.; Holzinger, A. Some aspects of the development of low-cost augmented reality learning environments as examples for future interfaces in technology enhanced learning. In Proceedings of the International Conference on Universal Access in Human-Computer Interaction, Beijing, China, 22–27 July 2007; Springer: Berlin/Heidelberg, Germany, 2007, pp. 728–737.
28. Vincenzi, D.A.; Valimont, B.; Macchiarella, N.; Opalenik, C.; Gangadharan, S.N.; Majoros, A.E. The effectiveness of cognitive elaboration using augmented reality as a training and learning paradigm. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Denver, CO, USA, 13–17 October 2003; SAGE Publications Sage CA: Los Angeles, CA, USA, 2003; Volume 47, pp. 2054–2058.
29. Binaee, K.; Diaz, G.J. Assessment of an augmented reality apparatus for the study of visually guided walking and obstacle crossing. *Behav. Res. Methods* **2019**, *51*, 523–531. [CrossRef]
30. Luchetti, A.; Parolin, E.; Butaslac, I.; Fujimoto, Y.; Kanbara, M.; Bosetti, P.; De Cecco, M.; Kato, H. Stepping over Obstacles with Augmented Reality based on Visual Exproprioception. In Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Recife, Brazil, 9–13 November 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 96–101.
31. Thatte, J.; Girod, B. Towards Perceptual Evaluation of Six Degrees of Freedom Virtual Reality Rendering from Stacked OmniStereo Representation. *Electron. Imaging* **2018**, *2018*, 352–1–352–6. [CrossRef]
32. Tatzgern, M.; Grasset, R.; Veas, E.; Kalkofen, D.; Seichter, H.; Schmalstieg, D. Exploring Real World Points of Interest. In Proceedings of the Pervasive Mobile Computing, St. Louis, MO, USA, 23–27 March 2015; Elsevier Science Publishers B. V.: Amsterdam, The Netherlands, 2015; Volume 18, pp. 55–70. doi: 10.1016/j.pmcj.2014.08.010. [CrossRef]
33. Tatzgern, M.; Grasset, R.; Kalkofen, D.; Schmalstieg, D. Transitional Augmented Reality navigation for live captured scenes. In Proceedings of the 2014 IEEE Virtual Reality (VR), Minneapolis, MN, USA, 29 March–2 April 2014; IEEE: Piscataway, NJ, USA, 2014; pp. 21–26. doi: 10.1109/VR.2014.6802045. [CrossRef]
34. Alonso, M.; Izaguirre, A.; Graña, M. Current research trends in robot grasping and bin picking. In Proceedings of the The 13th International Conference on Soft Computing Models in Industrial and Environmental Applications, San Sebastian, Spain, 6–8 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 367–376.
35. Billard, A.; Kragic, D. Trends and challenges in robot manipulation. *Science* **2019**, *364*, eaat8414. [CrossRef]
36. Rasouli, A.; Tsotsos, J.K. Autonomous vehicles that interact with pedestrians: A survey of theory and practice. *IEEE Trans. Intell. Transp. Syst.* **2019**, *21*, 900–918. [CrossRef]
37. Kang, H.J.; Shin, J.H.; Ponto, K. A Comparative Analysis of 3D User Interaction: How to Move Virtual Objects in Mixed Reality. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Atlanta, GA, USA, 22–26 March 2020; pp. 275–284.
38. Xiang, Y.; Schmidt, T.; Narayanan, V.; Fox, D. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. *arXiv* **2017**, arXiv:1711.00199. .
39. Sundermeyer, M.; Marton, Z.C.; Durner, M.; Triebel, R. Augmented autoencoders: Implicit 3d orientation learning for 6d object detection. *Int. J. Comput. Vis.* **2020**, *128*, 714–729. [CrossRef]
40. 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, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
41. Yang, Z.; Yu, X.; Yang, Y. Dsc-posenet: Learning 6dof object pose estimation via dual-scale consistency. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 3907–3916.
42. Wu, D.; Zhuang, Z.; Xiang, C.; Zou, W.; Li, X. 6D-VNet: End-To-End 6DoF Vehicle Pose Estimation From Monocular RGB Images. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019; pp. 1238–1247. doi: 10.1109/CVPRW.2019.00163. [CrossRef]
43. Lu, X.X. A review of solutions for perspective-n-point problem in camera pose estimation. *J. Phys. Conf. Ser.* **2018**, *1087*, 052009. [CrossRef]
44. Park, K.; Patten, T.; Vincze, M. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7668–7677.
45. Bukschat, Y.; Vetter, M. EfficientPose: An efficient, accurate and scalable end-to-end 6D multi object pose estimation approach. *arXiv* **2020**, arXiv:2011.043072020.
46. Tan, M.; Le, Q.E. Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
47. Yang, W.; Qian, Y.; Kämäräinen, J.K.; Cricri, F.; Fan, L. Object detection in equirectangular panorama. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2190–2195.

48. Zhao, P.; You, A.; Zhang, Y.; Liu, J.; Bian, K.; Tong, Y. Spherical criteria for fast and accurate 360 object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12959–12966.
49. Blair, C.; Walsh, C.; Best, P. Immersive 360° videos in health and social care education: A scoping review. *BMC Med. Educ.* **2021**, *21*, 1–28. [CrossRef]
50. Lampropoulos, G.; Barkoukis, V.; Burden, K.; Anastasiadis, T. 360-degree video in education: An overview and a comparative social media data analysis of the last decade. *Smart Learn. Environ.* **2021**, *8*, 1–24. [CrossRef]
51. Rambach, J.; Deng, C.; Pagani, A.; Stricker, D. Learning 6dof object poses from synthetic single channel images. In Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), Munich, Germany, 16–20 October 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 164–169.
52. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* **2016**, *17*, 2096–2130.
53. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2038–2041.
54. Shrivastava, A.; Pfister, T.; Tuzel, O.; Susskind, J.; Wang, W.; Webb, R. Learning from Simulated and Unsupervised Images through Adversarial Training. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2107–2116.
55. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. YOLACT++: Better Real-time Instance Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1108–1121. [CrossRef]
56. Ding, L.; Goshtasby, A. On the Canny edge detector. *Pattern Recognit.* **2001**, *34*, 721–725. [CrossRef]