3D Luggage Detection

Ziyu Wu, Heyuan Liu

March 14, 2025

Abstract

This project is aiming to enhance the luggage detection on a conveyor system by 3D information to accurately understand and reconstruct the shape of luggage, typically approximated as a cuboid with six facets. Building upon the previous solution of the 2D ResNet-based ALIX project, this project explores methodologies for identifying luggage using multiple calibrated views or even a single uncalibrated view. To achieve this, we are actively exploring several approaches. These include different architectures based on learning techniques integrating dense stereo alignment with epipolar constraints, and directly predicting cuboid shapes and their six key parameters (position and dimensions).

1 Background

The accurate detection and reconstruction of luggage on conveyor systems is a critical task in various industries, particularly in logistics, transportation, and security. With the growing demand for efficient baggage handling and inspection systems, automated solutions have become essential for streamlining operations, reducing human error, and improving throughput. Current solutions largely rely on 2D computer vision techniques, such as the ResNet-based[HZRS16] ALIX project[IDE23], which analyzes images of luggage captured by cameras. While these methods have proven effective for certain tasks, they are limited by their inability to fully capture the 3D geometry and spatial structure of luggage, which is essential for precise localization, volumetric analysis, and robust detection under diverse conditions.

To address these limitations, the integration of 3D information into luggage detection systems has emerged as a promising approach[HLB19]. Unlike 2D methods, 3D-based techniques can provide a more comprehensive understanding of the luggage shape, position, and orientation. This is particularly significant for conveyor systems, where luggage often overlaps, rotates, or appears partially occluded. By representing luggage as cuboids with six facets, 3D reconstruction methods can more accurately approximate the geometry and deliver better performance in real-world scenarios.[FDU12][YS19][DMBR16]

The motivation for this project stems from the need to overcome the challenges posed by traditional 2D methods and build a more reliable system that can handle complex environments. Leveraging 3D data not only enhances detection accuracy but also enables novel applications, such as volumetric analysis for luggage classification, anomaly detection, and improved object tracking.

2 Related Work

3D surface reconstruction techniques for luggage detection focus on external geometry rather than internal content analysis, distinguishing them from methods like X-ray computed tomography (CT), which primarily analyze internal structures. Methods such as stereo vision and structure-from-motion (SfM) have been widely used to recover 3D shapes from 2D images [Wu11]. However, these approaches often encounter challenges in dynamic conveyor environments, where luggage frequently overlaps, rotates, or appears partially occluded. Recent advancements in point cloud-based 3D object detection, particularly in autonomous driving and robotics, demonstrate the effectiveness of depth data for external surface analysis. The classic methods for this application are listed in the Table 1.

Deep learning has further improved 3D modeling through techniques like neural radiance fields (NeRF), enabling high-quality 3D reconstructions from sparse 2D views [MST+21]. While these

Category	Paper Title	Key Focus
3D Bounding Box	Deep3DBox [MAFK17]	3D bounding box estimation from monocular images
	GS3D [LOS+19]	Efficient 3D object detection using geometric models
	3D Object Localisation from Multi-view Image Detections [RCDB17]	Multi-view object localization
Bird's Eye View	Bird 3D Detection [SJS19]	Detection using bird's-eye view projections
Point Cloud-Based	P2VNet [YYHL23]	Learning representations from 3D point clouds
SfM Techniques	Multistage SfM: A Coarse-to-Fine Approach for 3D Reconstruction [SDN15]	Multi-stage structure-from-motion optimization
Structural Modeling	Robust 3D Reconstruction of Building Surfaces [WCC+20]	Reconstruction with structural and closed constraints with similar shapes of luggage

Table 1: Categorized Related Work

methods excel in generating detailed geometry, their high computational demands make them unsuitable for real-time conveyor belt scenarios. On the other hand, feature extraction methods such as SuperPoint[DMR18] and feature matching frameworks like SuperGlue[SDMR20] have shown potential in improving 3D reconstruction by enhancing feature quality, though their application to dynamic luggage detection remains underexplored.

Transfer learning offers a practical solution to address the lack of large labeled datasets in luggage detection. By fine-tuning pre-trained models on domain-specific tasks, such as those trained on large-scale datasets like ImageNet[DDS+09] or ShapeNet[CFG+15], researchers can reduce data requirements while maintaining strong performance. Efficient online transfer learning has further highlighted techniques for adapting 3D object classifiers to new domains with minimal computational costs, demonstrating their applicability in specialized scenarios.

This project leverages these advances to develop a robust framework for surface-level 3D modeling of luggage on conveyor systems. Unlike methods designed for internal analysis, our approach emphasizes external surface geometry, addressing challenges posed by dynamic and cluttered environments. By integrating dense stereo alignment, transfer learning, and geometric modeling, this work aims to deliver a scalable and efficient solution for real-world applications in logistics, tracking, and classification.

3 Technical Details

In this project, we restored the 3D size of luggage objects through multi-view photos, mainly referring to the method of the paper "3D Object Localisation from Multi-view Image Detections"[RCDB17]. The following is a description of the specific implementation process:

3.1 Data Format and Description

According to the reference method, the input data is organized into the following formats, as summarized in Table 2.

The dataset consists of photos of different luggage, each of which is photographed by a fixed camera array. The camera array contains four cameras that are located at different positions to capture multi-view images of the target object, Fig.1 shows an example. The intrinsic and extrinsic parameters of each camera group remain unchanged throughout the experiment, which means that for each baggage shot, the position and direction of the camera are fixed. The dataset contains about 1,000 different pieces of luggage, each piece of luggage corresponds to three sets of photos, each set contains four

photos (taken by four fixed cameras), totaling about 12,000 images. Such a data scale provides a sufficient training and testing basis for 3D reconstruction.

Figure 1: One set of data.

Data Type	Dimension	Description
Image Data	-	Multiple images containing target objects.
2D Bounding Box	$[n_{frames} \times (n_{objects} \times 4)]$	2D bounding box per object per frame, defined by the upper-left and lower-right corner coordinates.
Camera Intrinsics Matrix	$[3 \times 3]$	Intrinsic parameters of the camera, including focal length, optical center, and distortion coefficients.
Camera Extrinsics Matrix	$[n_{frames} \times (4 \times 3)]$	Camera pose per frame, represented by a rotation matrix and a translation vector.
Visibility Matrix	$[n_{frames} \times n_{objects}]$	Binary visibility indicator (1: visible, 0: invisible) for each object in each frame.

Table 2: Overview of Input Data Format

3.2 Camera Parameter Acquisition

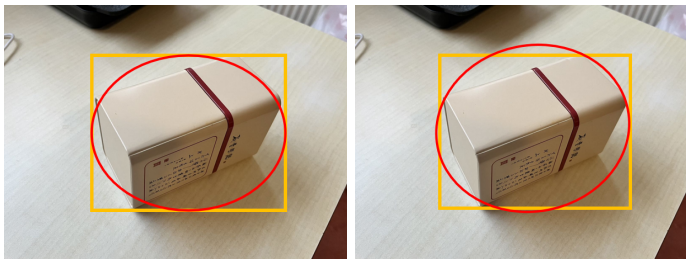
Since the data in our case does not contain the internal and external parameter information of the camera, additional steps are required to obtain these parameters. First, extract key points from multiple view images, and use algorithms such as SIFT, ORB, or AKAZE to detect local feature points in the image. Then use a matching algorithm (such as BFMatcher or SuperGlue) to match the feature points in multiple images and build a correspondence between multiple viewpoints. Finally, the camera's intrinsic parameter matrix (focal length, principal point coordinates, etc.) and external parameter information (position and posture) are obtained through the multi-view feature point matching relationship.

3.3 2D Bounding Box to Ellipse

In order to convert the bounding box data of 2D detection into a mathematical form suitable for geometric calculations, this method fits each bounding box as an ellipse. First, determine the center position of the ellipse through the center point of the bounding box, and use the width and height of the bounding box as the major and minor axes of the ellipse, respectively. Fig. 2a shows an example. Based on the above information, generate a mathematical description of the ellipse that satisfies the following constraints:

$$u^T \hat{C}_{if} u = 0, \text{ where } u = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \hat{C}_{if} \in \mathbb{R}^{3\times3} \tag{1}$$

Subsequently, the data needs to be preprocessed and optimized, since the tilt and orientation of the objects are not always consistent with the 2D bounding box. The processing methods include centralizing and normalizing the data to reduce the impact of different resolutions and scales on the calculation. The parameters of the ellipse are adjusted through optimization methods such as the least squares method to make it more consistent with the actual shape of the object. And further introduce tilt angle information to better describe the directionality of the target object, as shown in Fig.2b. This angle introduces the angle information in the camera view, object posture or detection results, providing more accurate initial geometric constraints for subsequent 3D reconstruction.



(a) Ellipse before pre-processing (b) Ellipse after pre-processing

Figure 2: 2D bounding box and ellipse.

3.4 Basic Concept of Dual Space

In order to simplify the mapping and solution from a two-dimensional ellipse to a three-dimensional ellipsoid, the concept of dual space is introduced in the method. Dual space expresses the geometric relationship between the ellipse and the ellipsoid in a linear form by describing the tangent line (in 2D) or tangent plane (in 3D). For 2D ellipse and 3D ellipsoid, their dual space representations are respectively Eq(2) and Eq(3). Among them, adj() represents the adjoint matrix operation, through which the description of the set of tangent lines of an ellipse or the set of tangent planes of an ellipsoid can be obtained.

$$\begin{gathered} \hat{C}_{if}^* = \text{adj}(\hat{C}_{if}) \in \mathbb{R}^{3\times3} \\ \text{where } u^T \hat{C}_{if} u = 0, \; u = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \; \hat{C}_{if} \in \mathbb{R}^{3\times3} \end{gathered} \tag{2}$$

$$\begin{gathered} Q_i^* = \text{adj}(Q_i) \in \mathbb{R}^{4\times4} \\ \text{where } x^T Q_i x = 0, \; x = \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \; Q_i \in \mathbb{R}^{4\times4} \end{gathered} \tag{3}$$

Furthermore, in dual space, the relationship between the dual space representation of a 2D ellipse and a 3D ellipsoid is Eq(4). $P_f \in \mathbb{R}^{3\times4}$ is the projection matrix, which is composed of the camera's intrinsic parameter matrix K_f and extrinsic parameter matrix $[R_f \| t_f]$.

$$\hat{C}_{if} = P_f Q_i^* P_f^T \tag{4}$$

3.5 Dual Space Solution Process

After determining the dual-space representation of 2D and 3D ellipsoids, this method uses mathematical methods to complete the reconstruction of 3D ellipsoids from multiple perspectives. To facilitate

calculation, the method vectorizes the 2D and 3D dual-space matrices. The vech() function is used to extract the lower triangular part of the symmetric matrix and flatten it into a vector:

$$c_{if}^* = \text{vech}(\hat{C}_{if}),\ v_i^* = \text{vech}(Q_i^*) \tag{5}$$

According to the linear relationship in the dual space, all the constraints of the perspective are superimposed to construct a linear equation system $G_f v_i^* = c_{if}^*$, where G_f is build from P_f. After stacking the constraints from different perspectives, we get the linear system $M_i w_i = 0$, where M_i is build from G_i and c_{if}^*, and w_i is build from v_i^* and scale factor. Then the constructed linear system is solved by the singular value decomposition (SVD) method to obtain the least squares solution of $M_i = U\Sigma V^T$. Finally, as an example shown in Fig.3, it should be noted that the paper mentions that due to the theory of multi-perspective geometry, at least three different angles are required to provide sufficient geometric constraints to obtain a definite solution.

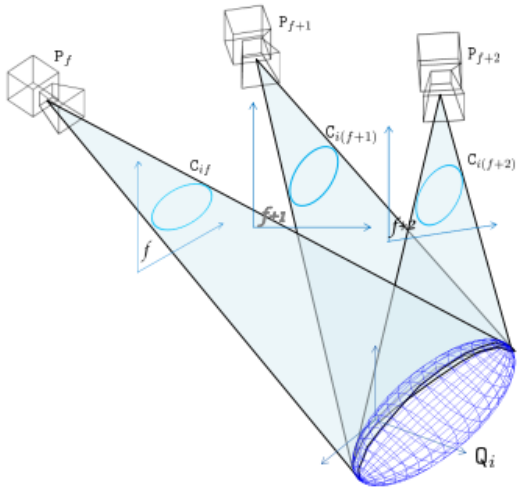


Figure 3: Multi-perspective recovery[RCDB17].

4 Achievements and Conclusion

4.1 Results

In the initial stage of the experiment, we first used the open source YOLOv11s[KH24] model to detect 2D bounding boxes. The results show that the model can perform well in the baggage detection task, accurately locate the baggage objects in each picture and generate bounding boxes. However, we encountered some problems when extracting the camera's internal and external parameters. Due to the low resolution of the images in the dataset, which is only 256 × 192, the number of detected feature points is very limited, which cannot meet the requirements of camera pose recovery (Fig.4).

Additionally, a substantial portion of the dataset proved unsuitable for camera reconstruction, as most images did not capture the entire object, further restricting the available visual information for multi-view geometry processing.

To address this problem, we tried to preprocess the images. First, we converted the color images into grayscale images and supersampled them to increase the resolution. This method increased the number of feature points to a certain extent, but the matching effect was still poor, resulting in the inability to recover the camera pose (Fig.6d). Next, we further detected the 2D bounding box on the supersampled image, cropped the bounding box area, and used semantic segmentation technology to remove the background, hoping to reduce background interference in this way. However, this method still failed to significantly improve the effect of feature point matching (Fig.6c).

To further improve the performance of feature point detection and matching, we tried a variety of open source models and tools, including Colmap[SF16, SZPF16], VisualSFM[Wu11], PoseNet[KGC15], SuperGlue[SDMR20], and Meshroom[GGC+21]. These tools usually perform very well on public

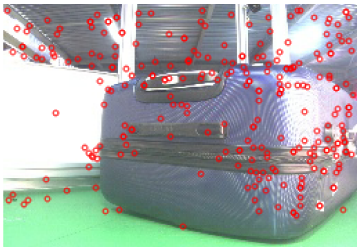
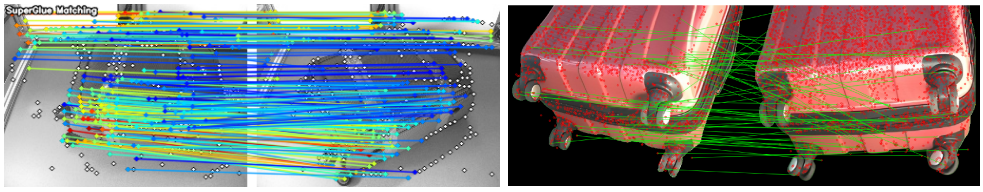


Figure 4: Feature points on low resolution data.

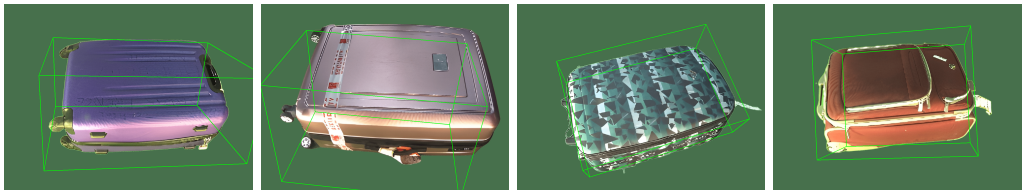


(a) Feature matching after supersampled (b) Feature matching after segmentation

Figure 5: 2D bounding box and ellipse.

datasets, but they still failed to achieve ideal results on our experimental data. The number of feature points and matching accuracy were still insufficient, which ultimately led to the failure of camera pose recovery.

Through multiple image experiments, we observed that suitcases with textured surfaces consistently yielded better performance in aligning the 3D bounding box with the object in images, as showin in Fig.6. This finding suggests that an increased number of distinctive feature points improves the accuracy of matching.



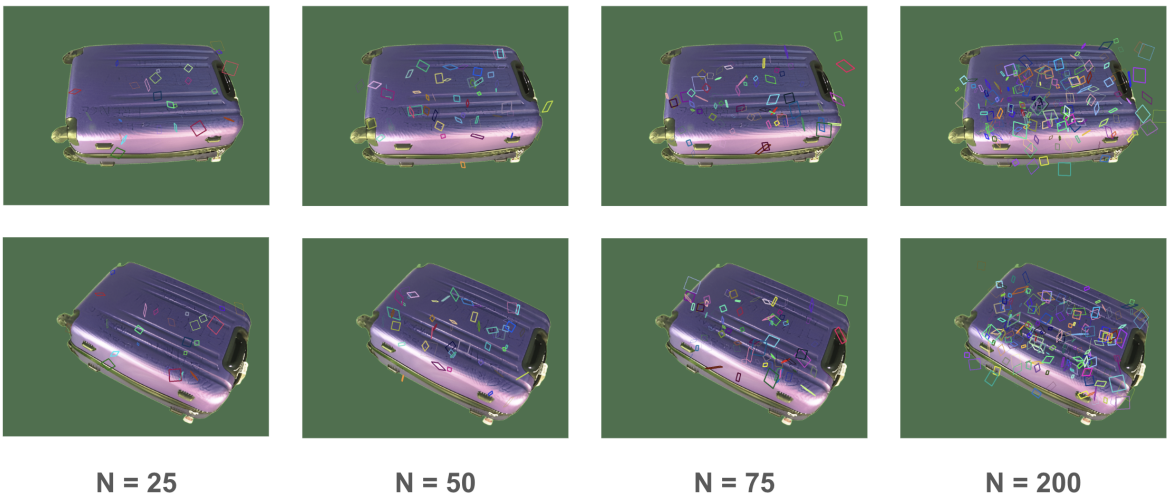
(a) Normal Surface (b) Normal Surface (c) Textured Surface (d) Normal Surface

Figure 6: Textured surface has a better performance

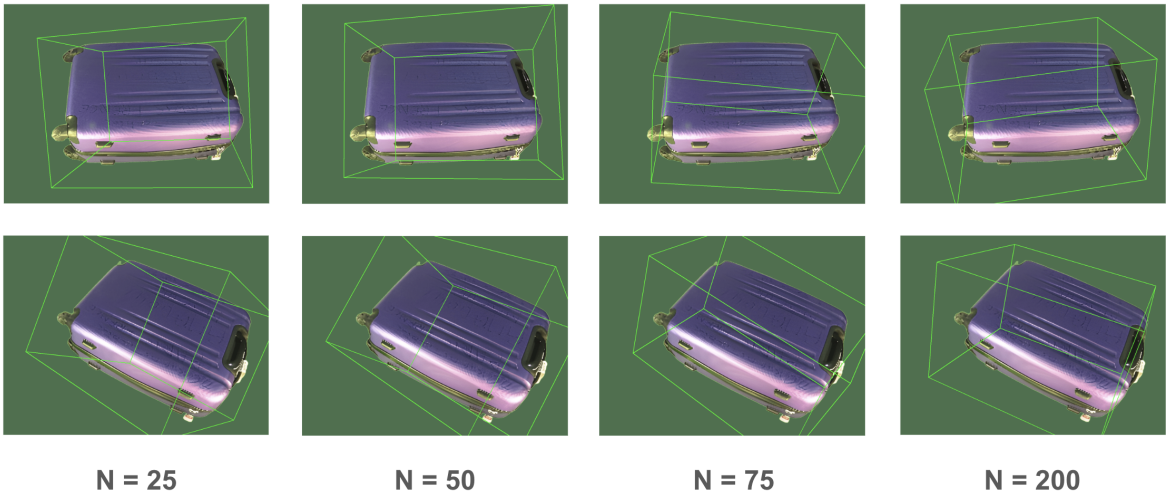
Motivated by this observation, to enhance feature extraction and improve 3D box estimation, we first introduced geometric patterns and textures directly onto the 3D box model. This modification enriched the surface details and artificially increased the number of feature points, enabling a constraint-based approach to improve box identification. By increasing the feature density, we aimed to achieve better alignment between the 3D bounding box and the object in images, ultimately enhancing the accuracy of matching and reconstruction. Initial experiments demonstrated that adding a fixed number of shapes to the 3D scene (Fig.7a) improved box estimation. However, despite these enhancements, mismatches still persisted (Fig.7b), indicating that a more refined approach was needed.

To further improve 3D reconstruction accuracy, we adopted an iterative optimization process (as illustrated in Fig.8). Initially, we performed a baseline 3D reconstruction using the available image data, but due to the lack of distinctive surface details on many luggage models, feature extraction and matching remained suboptimal, leading to inaccuracies in the reconstructed geometry.

To mitigate this issue, we systematically refined the augmentation by iteratively adjusting the geometric patterns and textures applied to the 3D model. After introducing additional textures, we projected the modified 3D model back onto the camera plane, generating updated image projections that preserved scene consistency while incorporating richer feature information. With these enhanced images, we performed more rounds of 3D reconstruction to assess improvements in alignment and camera pose estimation.



(a) Different Numbers of Shapes added



(b) Box estimation after augmentation

Figure 7: Augmented Box Estimation with example 1

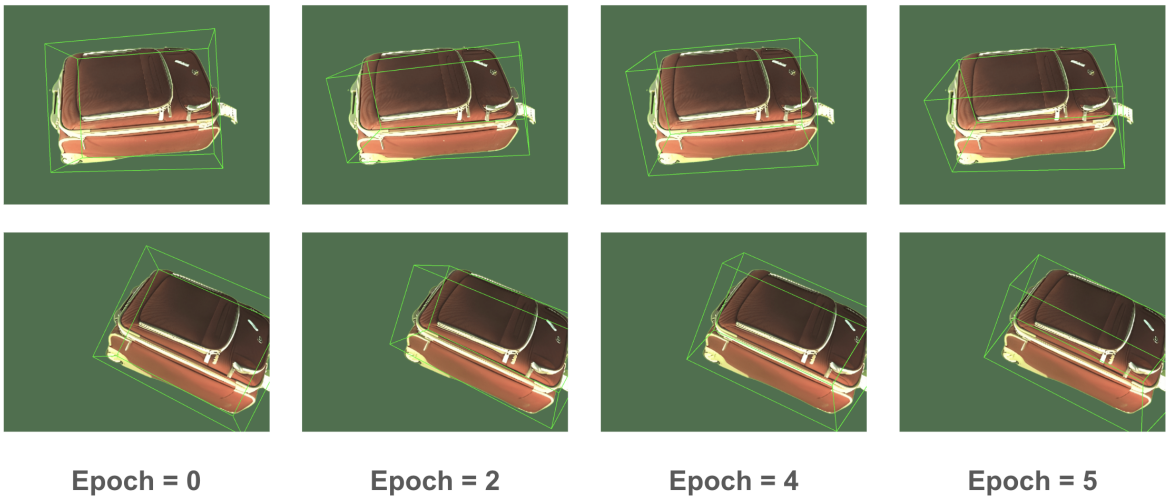


Figure 8: Iterations on adding textures with N=25 shapes

4.2 Conclusion

This project explores the integration of 3D information into luggage detection and reconstruction systems, addressing the limitations of traditional 2D approaches. By leveraging geometric modeling and feature extraction techniques, we have demonstrated that enhancing surface textures significantly improves feature matching and 3D reconstruction accuracy. Through extensive experimentation, we observed that smooth luggage surfaces posed a major challenge for feature detection, leading to suboptimal camera pose estimation and reconstruction errors. To mitigate this, we systematically introduced geometric patterns onto 3D models, iteratively refining the augmentation process to optimize feature density and alignment.

Our results indicate that while adding textures improves feature extraction and 3D box estimation, challenges such as feature mismatches and reconstruction inconsistencies remain. The iterative refinement process has shown promise in reducing these errors, but further optimization is required to achieve a robust and scalable solution. Future work will focus on automating the texture enhancement process, improving feature detection algorithms, and integrating additional deep learning models to refine pose estimation and geometric consistency. Ultimately, this study highlights the potential of 3D-informed approaches in enhancing luggage detection, paving the way for more accurate, efficient, and automated conveyor belt monitoring systems in logistics and security applications.

References

[CFG+15] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DMBR16] Debidatta Dwibedi, Tomasz Malisiewicz, Vijay Badrinarayanan, and Andrew Rabinovich. Deep cuboid detection: Beyond 2d bounding boxes. *arXiv preprint arXiv:1611.10010*, 2016.

[DMR18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[FDU12] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. *Advances in neural information processing systems*, 25, 2012.

[GGC+21] Carsten Griwodz, Simone Gasparini, Lilian Calvet, Pierre Gurdjos, Fabien Castan, Benoit Maujean, Gregoire De Lillo, and Yann Lanthony. Alicevision Meshroom: An open-source 3D reconstruction pipeline. In *Proceedings of the 12th ACM Multimedia Systems Conference - MMSys '21*. ACM Press, 2021.

[HLB19] Xian-Feng Han, Hamid Laga, and Mohammed Bennamoun. Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1578–1604, 2019.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[IDE23] IDEMIA. ALIX: Automating lost luggage identification processes using artificial intelligence. *IDEMIA Press Release*, March 2023. Accessed: 2024-12-17.

[KGC15] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*, pages 2938–2946, 2015.

[KH24] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

[LOS+19] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1019–1028, 2019.

[MAFK17] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017.

[MST+21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[RCDB17] Cosimo Rubino, Marco Crocco, and Alessio Del Bue. 3d object localisation from multi-view image detections. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1281–1294, 2017.

[SDMR20] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

[SDN15] Rajvi Shah, Aditya Deshpande, and PJ Narayanan. Multistage sfm: A coarse-to-fine approach for 3d reconstruction. *arXiv preprint arXiv:1512.06235*, 2015.

[SF16] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[SJS19] Siddharth Srivastava, Frederic Jurie, and Gaurav Sharma. Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4504–4511. IEEE, 2019.

[SZPF16] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[WCC+20] Senyuan Wang, Guorong Cai, Ming Cheng, José Marcato Junior, Shangfeng Huang, Zongyue Wang, Songzhi Su, and Jonathan Li. Robust 3d reconstruction of building surfaces from point clouds based on structural and closed constraints. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170:29–44, 2020.

[Wu11] Changchang Wu. Visualsfm: A visual structure from motion system. *http://www. cs. washington. edu/homes/ccwu/vsfm*, 2011.

[YS19] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35(4):925–938, 2019.

[YYHL23] Jun Yu, Wenbin Yin, Zhiyi Hu, and Yabin Liu. 3d reconstruction for multi-view objects. *Computers and Electrical Engineering*, 106:108567, 2023.