6-DoF Object Pose Estimation

Subjects: Computer Science, Artificial Intelligence
Contributor: Zihang Wang, Xueying Sun, Hao Wei, Qing Ma, Qiang Zhang

Accurately estimating the six-degree-of-freedom (6-DoF) of objects is a critical task in various applications, including robotics, autonomous driving, and virtual reality. For instance, the precise estimation of spatial coordinates and rotational orientation of an object is essential for robotic tasks such as manipulation, navigation, and assembly.

Keywords: object pose estimation ; six-degree-of-freedom ; 6-DoF

1. Introduction

Accurately estimating the six-degree-of-freedom (6-DoF) of objects is a critical task in various applications, including robotics, autonomous driving, and virtual reality. For instance, the precise estimation of spatial coordinates and rotational orientation of an object is essential for robotic tasks such as manipulation, navigation, and assembly. However, achieving robustness in 6-DoF detection remains a challenging problem. In real-world applications, numerous object types exhibit significant occlusions and variations in lighting conditions. Due to the increasing reliability of new RGB-D image sensors, the 6-DoF detection of visual targets based on multi-source image information is flourishing. Researchers have explored a number of ways [1][2][3] to fuse RGB image data and depth image data to guide 6-DoF detection of visual targets with impressive accuracy. Different research teams are employing various framework approaches to investigate solutions for the 6DoF pose estimation problem. Some focus on the overall algorithmic framework, while others delve into efficient feature extraction.

Regarding the problem of object pose estimation, previous approaches predominantly employed adaptive matrices to tackle this issue. However, with the rise of convolutional neural networks (CNN) and transformers, deep learning (DL) based methods are used to solve the 6-DoF estimation problem. There are two main types of DL-based frameworks for 6D attitude estimation of objects: end-to-end architectures [4][5] and two-stage segmentation-pose regression architectures [6][7]. End-to-end models integrate multiple stages of visual processing steps into a single model; therefore, their networks are less complex and computationally intensive. A single network processes pixel information from the image to deduce the region where the candidate target is located and its corresponding 6DoF pose information. The internal structure and decision-making process of this neural network are more hidden, less interpretable, and more difficult to train. On the other hand, the two-stage segmentation-pose regression architecture first segments the visual target from the scene and then obtains the pose of the visual target in the scene by regression. This method is able to focus on the visual target being detected and exclude interference from the background, resulting in more reliable results.

In the process of 6DoF pose estimation through image features, there have been numerous prior efforts. Some have employed manually designed features (such as SIFT) to extract object characteristics for subsequent pose regression. However, the limited quantity of manually designed features might lead to failures in pose regression. Depth images provide dense features, yet enhancing the robustness of these depth features remains an unsolved challenge. Solely relying on RGB or depth information addresses only one facet of the problem. Thus, the approach leverages the fusion of RGB-D data to accomplish the task. Prior research has made significant strides in exploring the fusion of RGB and depth images. A multitude of studies have delved deeply into various techniques and algorithms aiming to effectively exploit the complementary information these modalities provide. However, despite these commendable efforts, achieving seamless integration between RGB and depth images remains an ongoing and formidable challenge. Existing methods often grapple with the intricate task of synchronizing the two modalities accurately, resulting in less than optimal fusion outcomes. Moreover, inherent differences in intrinsic features between RGB and depth data, including variations in lighting conditions and occlusions, further amplify the complexity of the fusion process. As such, continuous research and innovation are urgently needed to elevate the fusion of RGB and depth images in target pose detection to new heights.

2. Enhancing 6-DoF Object Pose Estimation

2.1. Feature Representation

In vision tasks, the representation of image features plays a crucial role in various applications, including visual target recognition and detection. In the context of target pose estimation, it is essential for the features of visual targets to exhibit robustness against translation, rotation, and scaling. Additionally, these features should possess local descriptive capabilities and resistance to noise.

In previous studies, researchers have utilized image feature matching to detect the position of visual targets. The pose of the target can be obtained by solving the 2D-to-3D PnP problem. Artificially designed features such as SIFT [8][9], SURF [10], DAISY [11], ORB [12], BRIEF [13], BRISK [14], and FREAK [15] have demonstrated robustness against occlusion and scale-scaling issues. These descriptors have been widely adopted in models for target position detection. Similarly, 3D local features such as PFH [16][17][18], FPFH [19], SHOT [20], C-SHOT [21], and RSD [22] can effectively extract features and detect the position of targets in 3D point clouds. Recently, machine learning based feature descriptor algorithms [23][24] are receiving more and more attention in the field of image matching. These methods employ PCA [25], random trees [26], random fern [27], and boosting [28] algorithms to achieve more robust features than hand-designed features.

However, in cases where the surface of the visual target is smooth and lacks texture, the extraction of manually designed feature points is often limited in number. This limitation adversely affects the reliability of object pose estimation. Furthermore, the high apparent similarity among visual targets also poses challenges in accurately estimating the positional attitude of the detected target.

In addition to manually designed features, there are supervised learning-based feature description methods such as triplet CNN descriptor [29], LIFT [30], L2-net [31], HardNet [32], GeoDesc [33]. For the recognition of textureless objects, global features can be implemented by utilizing image gradients or surface normals as shape attributes. Among these, template-based global features aim to identify the region in the observed image that bears the closest resemblance to the object template. Some commonly employed template-based algorithms include Line-MOD [34] and DTT-OPT [35]. In recent years, novel 3D deep learning methods have emerged, such as OctNet [36], PointNet [37], PointNet++ [38], and MeshNet [39]. These methods are capable of extracting distinctive deep representations through learning and can be employed for 3D object recognition or retrieval.

2.2. Two-Stage or Single-Shot Approach

In the realm of object 6D pose estimation frameworks, two main types can be identified: end-to-end architectures and two-stage segmentation 6-DoF regression architectures.

In the field of object detection, notable end-to-end frameworks like YOLO [40] and SSD [41] have emerged. These frameworks have been extended to address the challenge of target pose detection. Poirson et al. [42] proposed an end-to-end object and pose detection architecture based on SSD, treating pose estimation as a classification problem using RGB images. Another extension, SSD-6D [43], utilizes multi-scale features to regress bounding boxes and classify pose into discrete viewpoints. Yisheng He et al. [44] introduced PVN3D, a method based on a deep 3D Hough voting network that fuses appearance and geometric information from RGB-D images.

Two-stage architectures segment the visual target and estimate pose through regression. For example, in [45], pose estimation was treated as a classification problem using the 2D bounding box. Mousavian et al. [46] utilized a VGG backbone to classify pose based on the 2D bounding box and regress the offset. Nuno Pereira et al. [7] proposed MaskedFusion, a two-stage network that employed an encoder–decoder architecture for image segmentation and utilized fusion with RGB-D data for pose estimation and refinement. This two-stage neural network effectively leverages the rich semantic information provided by RGB images and exhibits good decoupling, allowing for convenient code replacement when improvements are required for a specific stage algorithm. Additionally, this design helps reduce training costs.

However, the MaskedFusion method employed in the first stage solely relies on RGB image information, which often leads to insufficient and inaccurate semantic information in low-light and low-texture scenarios. This results in issues such as blurry edges and erroneous segmentation in the Mask image of the segmentation network during practical scene applications.

2.3. Single Modality or Multi-Modality Fusion

2.3.1. RGB Single Modal Based Object Pose Estimation

For visual target position detection, RGB images have traditionally been used as the primary data source. Feature matching techniques are commonly employed for localizing target positions within 2D images. PoseCNN [47] utilizes a convolutional neural network and Hough voting to estimate the target's pose. PvNet [48] extracts keypoints from RGB images and employs a vector field representation for localization. Hu et al. [49] proposed a segmentation-driven framework that uses a CNN to extract features from RGB image and assigns target category labels to virtual meshes. The ROPE framework [50] incorporates holistic pose representation learning and dynamic amplification for accurate and efficient pose estimation. SilhoNet [51] also predicts object poses using a pipeline with a convolutional neural network. Zhang et al. [52] proposed an end-to-end deep learning architecture for object detection and pose recovery from single RGB modal data. Aing et al. [53] introduced informative features and techniques for segmentation and pose estimation.

Although image-based methods have achieved promising results in 6-DoF estimation, their performance tends to degrade when dealing with textureless and occluded scenarios.

2.3.2. 3D Cloud or Depth Image Based Object Pose Estimation

Recovering the position of a visual target from 3D point cloud or depth image data is also a common method. The RGM method [54] introduces deep graph matching for point cloud registration, leveraging correspondences and graph structure to address outliers. This approach replaces explicit feature matching and RANSAC with an attention mechanism, enabling an end-to-end framework for direct prediction of correspondence sets. Rigid transformations can be estimated directly from the predicted correspondences without additional post-processing. The BUFFER method [55] enhances computational efficiency by predicting key points and improves feature representation by estimating their orientation. It utilizes a patch-wise embedder with a lightweight local feature learner for efficient and versatile piecewise features. The ICG framework [56] presents a probabilistic tracker that incorporates region and depth information, relying solely on object geometry

Nonetheless, the point cloud data inherently exhibits sparsity and lacks sufficient texture information, which poses limitations to the performance of these methods. Consequently, the incorporation of RGB image information represents a crucial enhancement to enhance the accuracy and effectiveness of the position estimation.

2.3.3. Multi-Modal Data Based Object Pose Estimation

In the field of target position detection, the fusion of information from multiple sensors has emerged as a cutting-edge research area for accurate position detection. Zhang et al. [57] proposed a hybrid Transformer-CNN method for 2-DoF object pose detection. They further proposed a bilateral neural network architecture [58] for RGB and depth image fusion and achieved promising results. In 6-DoF pose detection area, Wang et al. [6] introduced the DenseFusion framework for precise 6-DoF pose estimation using two data sources and a dense fusion network. MaskedFusion [7] achieved superior performance by incorporating object masking in a pipeline. Se(3)-TrackNet [59] presented a data-driven optimization approach for long-term 6D pose tracking. PVN3D [44] adopted a keypoint-based approach for robust 6DoF object pose estimation from a single RGBD image. FFB6D [5] introduced a bi-directional fusion network for 6D bit-pose estimation, exploiting the complementary nature of RGB and depth images. The ICG+ [60] algorithm incorporated additional texture patterns for flexible multi-camera information fusion. However, existing methods still face challenges in extracting feature information from RGB-D data.

References

1. Huang, X.; Mei, G.; Zhang, J.; Abbas, R. A Comprehensive Survey on Point Cloud Registration. arXiv 2021, arXiv:2103.02690.
2. Zhu, Y.; Li, M.; Yao, W.; Chen, C. A Review of 6D Object Pose Estimation. In Proceedings of the 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 17–19 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1647–1655.
3. Du, G.; Wang, K.; Lian, S.; Zhao, K. Vision-Based Robotic Grasping from Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review. Artif. Intell. Rev. 2021, 54, 1677–1734.
4. Amini, A.; Periyasamy, A.S.; Behnke, S. T6D-Direct: Transformers for Multi-Object 6D Pose Direct Regression. In DAGM German Conference on Pattern Recognition; Springer: Cham, Switzerland, 2021; Volume 13024, pp. 530–544.
5. He, Y.; Huang, H.; Fan, H.; Chen, Q.; Sun, J. FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 3003–3013.

6. Wang, C.; Xu, D.; Zhu, Y.; Martín-Martín, R.; Lu, C.; Fei-Fei, L.; Savarese, S. Densefusion: 6d object pose estimation by iterative dense fusion. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 3343–3352.
7. Pereira, N.; Alexandre, L.A. MaskedFusion: Mask-Based 6D Object Pose Estimation. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 14–17 December 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 71–78.
8. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
9. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 2004, 60, 91–110.
10. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). Comput. Vis. Image Underst. 2008, 110, 346–359.
11. Tola, E.; Lepetit, V.; Fua, P. DAISY: An Efficient Dense Descriptor Applied to Wide-Baseline Stereo. IEEE Trans. Pattern Anal. Mach. Intell. 2010, 32, 815–830.
12. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
13. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary Robust Independent Elementary Features. In Proceedings of the European Conference on Computer Vision (ECCV), Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 778–792.
14. Leutenegger, S.; Chli, M.; Siegwart, R.Y. BRISK: Binary Robust Invariant Scalable Keypoints. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2548–2555.
15. Alahi, A.; Ortiz, R.; Vandergheynst, P. FREAK: Fast Retina Keypoint. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 510–517.
16. Rusu, R.B.; Blodow, N.; Marton, Z.C.; Beetz, M. Aligning Point Cloud Views Using Persistent Feature Histograms. In Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nice, France, 22–26 September 2008; pp. 3384–3391.
17. Rusu, R.B.; Marton, Z.C.; Blodow, N.; Beetz, M. Learning Informative Point Classes for the Acquisition of Object Model Maps. In Proceedings of the Robotics and Vision 10th International Conference on Control, Automation, Hanoi, Vietnam, 17–20 December 2008; pp. 643–650.
18. Rusu, R.B.; Marton, Z.C.; Blodow, N.; Beetz, M. Persistent Point Feature Histograms for 3D Point Clouds. In Proceedings of the 10th International Conference on Intelligent Autonomous Systems (IAS-10), Baden-Baden, Germany, 23–25 July 2008; pp. 119–128.
19. Rusu, R.B.; Blodow, N.; Beetz, M. Fast Point Feature Histograms (FPFH) for 3D Registration. In Proceedings of the 2009 IEEE International Conference on Robotics and Automation, Kobe, Japan, 12–17 May 2009; pp. 3212–3217.
20. Salti, S.; Tombari, F.; Di Stefano, L. SHOT: Unique Signatures of Histograms for Surface and Texture Description. Comput. Vis. Image Underst. 2014, 125, 251–264.
21. Fischler, M.A.; Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Commun. ACM 1981, 24, 381–395.
22. Marton, Z.-C.; Pangercic, D.; Blodow, N.; Kleinehellefort, J.; Beetz, M. General 3D Modelling of Novel Objects from a Single View. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 3700–3705.
23. Brown, M.; Hua, G.; Winder, S. Discriminative Learning of Local Image Descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 2011, 33, 43–57.
24. Snavely, N.; Seitz, S.M.; Szeliski, R. Modeling the World from Internet Photo Collections. Int. J. Comput. Vis. 2008, 80, 189–210.
25. Ke, Y.; Sukthankar, R. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, Washington, DC, USA, 27 June–2 July 2004; Volume 2, pp. II-506–II-513.
26. Lepetit, V.; Fua, P. Keypoint Recognition Using Randomized Trees. IEEE Trans. Pattern Anal. Mach. Intell. 2006, 28, 1465–1479.
27. Ozuysal, M.; Calonder, M.; Lepetit, V.; Fua, P. Fast Keypoint Recognition Using Random Ferns. IEEE Trans. Pattern Anal. Mach. Intell. 2010, 32, 448–461.

28. Chen, L.; Rottensteiner, F.; Heipke, C. Learning Image Descriptors for Matching Based on Haar Features. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XL-3 2014, 40, 61–66.
29. Kumar, B.G.V.; Carneiro, G.; Reid, I. Learning Local Image Descriptors with Deep Siamese and Triplet Convolutional Networks by Minimising Global Loss Functions. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 5385–5394.
30. Yi, K.M.; Trulls, E.; Lepetit, V.; Fua, P. LIFT: Learned Invariant Feature Transform. In Proceedings of the Computer Vision (ECCV) 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 467–483.
31. Tian, Y.; Fan, B.; Wu, F. L2-Net: Deep Learning of Discriminative Patch Descriptor in Euclidean Space. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 661–669.
32. Mishchuk, A.; Mishkin, D.; Radenovic, F.; Matas, J. Working Hard to Know Your Neighbor's Margins: Local Descriptor Learning Loss. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Dutchess County, NY, USA, 2017; Volume 30.
33. Luo, Z.; Shen, T.; Zhou, L.; Zhu, S.; Zhang, R.; Yao, Y.; Fang, T.; Quan, L. GeoDesc: Learning Local Descriptors by Integrating Geometry Constraints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 168–183.
34. Hinterstoisser, S.; Holzer, S.; Cagniart, C.; Ilic, S.; Konolige, K.; Navab, N.; Lepetit, V. Multimodal Templates for Real-Time Detection of Texture-Less Objects in Heavily Cluttered Scenes. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 858–865.
35. Rios-Cabrera, R.; Tuytelaars, T. Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 2048–2055.
36. Riegler, G.; Osman Ulusoy, A.; Geiger, A. OctNet: Learning Deep 3D Representations at High Resolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 3577–3586.
37. Charles, R.Q.; Su, H.; Kaichun, M.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 77–85.
38. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Dutchess County, NY, USA, 2017; Volume 30.
39. Feng, Y.; Feng, Y.; You, H.; Zhao, X.; Gao, Y. MeshNet: Mesh Neural Network for 3D Shape Representation. Proc. AAAI Conf. Artif. Intell. 2019, 33, 8279–8286.
40. 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, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 779–788.
41. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part I 14. pp. 21–37.
42. 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 2016 Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 676–684.
43. 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 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1530–1538.
44. He, Y.; Sun, W.; Huang, H.; Liu, J.; Fan, H.; Sun, J. PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11629–11638.
45. Tulsiani, S.; Malik, J. Viewpoints and Keypoints. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; IEEE: Piscataway, NJ, USA; pp. 1510–1519.
46. Mousavian, A.; Anguelov, D.; Flynn, J.; Košecká, J. 3D Bounding Box Estimation Using Deep Learning and Geometry. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA,

21–26 July 2017; pp. 5632–5640.

47. 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.
48. Peng, S.; Liu, Y.; Huang, Q.; Bao, H.; Zhou, X. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. IEEE Trans. Pattern Anal. Mach. Intell. 2022, 44, 3212–3223.
49. Hu, Y.; Hugonot, J.; Fua, P.; Salzmann, M. Segmentation-Driven 6D Object Pose Estimation. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA; pp. 3380–3389.
50. Chen, B.; Chin, T.-J.; Klimavicius, M. Occlusion-Robust Object Pose Estimation with Holistic Representation. In Proceedings of the 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2022; pp. 2223–2233.
51. Billings, G.; Johnson-Roberson, M. SilhoNet: An RGB Method for 6D Object Pose Estimation. IEEE Robot. Autom. Lett. 2019, 4, 3727–3734.
52. Zhang, X.; Jiang, Z.; Zhang, H. Real-Time 6D Pose Estimation from a Single RGB Image. Image Vis. Comput. 2019, 89, 1–11.
53. Aing, L.; Lie, W.-N.; Lin, G.-S. Faster and Finer Pose Estimation for Multiple Instance Objects in a Single RGB Image. Image Vis. Comput. 2023, 130, 104618.
54. Fu, K.; Liu, S.; Luo, X.; Wang, M. Robust Point Cloud Registration Framework Based on Deep Graph Matching. IEEE Trans. Pattern Anal. Mach. Intell. 2023, 45, 6183–6195.
55. Ao, S.; Hu, Q.; Wang, H.; Xu, K.; Guo, Y. BUFFER: Balancing Accuracy, Efficiency, and Generalizability in Point Cloud Registration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1255–1264.
56. Stoiber, M.; Sundermeyer, M.; Triebel, R. Iterative Corresponding Geometry: Fusing Region and Depth for Highly Efficient 3D Tracking of Textureless Objects. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 6845–6855.
57. Zhang, Q.; Zhu, J.; Sun, X.; Liu, M. HTC-Grasp: A Hybrid Transformer-CNN Architecture for Robotic Grasp Detection. Electronics 2023, 12, 1505.
58. Zhang, Q.; Sun, X. Bilateral Cross-Modal Fusion Network for Robot Grasp Detection. Sensors 2023, 23, 3340.
59. Wen, B.; Mitash, C.; Ren, B.; Bekris, K.E. Se(3)-TrackNet: Data-Driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 10367–10373.
60. Stoiber, M.; Elsayed, M.; Reichert, A.E.; Steidle, F.; Lee, D.; Triebel, R. Fusing Visual Appearance and Geometry for Multi-Modality 6DoF Object Tracking. arXiv 2023, arXiv:2302.11458.