A SURVEY FOR 3D COMPUTER VISION IN ROBOTICS

Zhiyu Liu, Guang Yin, JiaWei Zhang
3D Vision
University of Illinois Urbana Champaign
Champaign, Illinois
{zhiyul6, guangy2, jiaweiz9}@illinois.edu

ABSTRACT

With the rapid advancement of the AI industry, the robotics industry is also thriving, becoming an integral part of various sectors of production and daily life. Robots are increasingly entering the public domain, offering enhanced automation and convenience. As these robots continue to evolve, they rely heavily on sophisticated perception systems to navigate and interact with environments.

In this research survey, we mainly focus on key advancements in 3D vision technologies that enable robots to perceive, understand, and operate effectively in complex and dynamic environments. We categorize the discussion into three major areas: Perception and Environment Modeling, Motion Perception, and Visual Decision Making. These areas encompass crucial topics like Simultaneous Localization and Mapping (SLAM), 3D reconstruction, pose estimation, and visual policies, which together empower robots to achieve greater autonomy. Through an analysis of current methodologies and the integration of cutting-edge techniques, we highlight both the capabilities and limitations of these technologies in real-world applications.

***K*eywords** SLAM · 3D Reconstruction · Visual Policy · Dynamic Model · Motion Perception

1 Introduction

In recent years, integrating artificial intelligence with robotics has significantly transformed industries ranging from manufacturing to healthcare. Robots used in logistics, service, and household tasks, have become increasingly common, driving the need for more intelligent robots. To achieve a higher level of intelligence, robots have to perceive and understand their three-dimensional surroundings.

However, traditional vision-based robotic systems, which often rely on two-dimensional visual inputs, face significant limitations in dynamic and unstructured environments. Accurate 3D perception is crucial for robots to effectively navigate, manipulate objects, and interact with the real world.

This survey provides a comprehensive review of the key components that enable 3D vision in robotics, including Perception and Environment Modeling, Motion Perception, and Visual Decision Making. We review recent advancements, highlight the limitations of existing approaches, and propose potential future directions to address the challenges of real-time performance and dynamic environments. Through this survey, we aim to offer insights into how cutting-edge techniques such as NeRF and 3D Gaussian Splatting are reshaping the landscape of robotic perception.

2 Environment Perception and Modeling

Over the past decade, significant advancements have been made in the field of 3D vision for robotics, particularly with the development of methods like Simultaneous Localization and Mapping (SLAM), 3D reconstruction, and semantic segmentation.

2.1 Visual SLAM

2.1.1 V-SLAM Architecture

V-SLAM framework is composed of sequential steps, which are shown in Fig1, including data acquisition and initialization, localization, mapping, and loop closure.

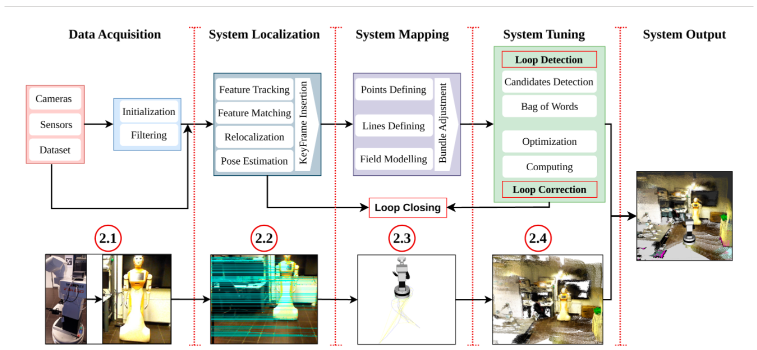

Figure 1: V-SLAM architecture: an overview of the four core components. Image credits:[1]

2.1.2 SOTA V-SLAM Methods

V-SLAM plays an important role within the robotics field and research. The landscape of V-SLAM is composed of a variety of methodologies, which can be divided into three categories, namely, pure visual SLAM, visual-inertial SLAM, and RGB-D SLAM.[1]

Pure Visual SLAM uses monocular, RGB-D, and stereo cameras to scan the environment, helping robots map unfamiliar areas easily. ORB-SLAM series: ORB-SLAM1[2] categorized to be only visual, while ORB-SLAM2[3] expands to both only-visual and RGB-D SLAM. Furthermore, ORB-SLAM3[4] furthers its classification to include all three categories: only-visual, visual-inertial, and RGB-D SLAM.

Visual-Inertial SLAM is a technique that combines the capabilities of visual sensors, such as stereo cameras, and inertial measurement sensors (IMUs) to achieve its SLAM objectives and operations. OKVIS-SLAM[5] uses image retrieval to connect keyframes in the SLAM pose-graph, aided by the pose estimator for locations beyond the optimization window of visual-inertial odometry. VINS Mono-SLAM[6] combines visual and inertial data to enhance accuracy and ensure precise functionality of robot operations.

RGB-D SLAM is an innovative approach that integrates RGB-D cameras with depth sensors to estimate and to build models of the environment. DTAM-SLAM[7] provides robust six degrees of freedom tracking and facilitates efficient environmental modeling for robotic systems. Since DTAM-SLAM is slightly dynamic with lights, it is accurate to operate in high and strong illumination fields.

2.2 3D Reconstruction

To meet the needs of environment modeling, collision detection, path planning, etc., researchers have devoted themselves to developing methods and algorithms for robots to autonomously construct increasingly highly accurate scenes.[8]

2.2.1 NeRF

NeRF[9] is introduced by Mildenhall et al. in 2020. It is an implicit, continuous volumetric representation, setting a new standard for novel view synthesis.

NeRF synthesizes images by sampling 5D coordinates(location(x, y, z) and viewing direction(pitch, yaw)) along camera rays, feeding those locations into an MLP to produce a color and volume density, and using volume rendering techniques to composite these values into an image. This rendering function is differentiable, so NeRF can optimize scene representation by minimizing the residual between synthesized and ground truth observed images.

While NeRF achieved success, challenges like slow training/rendering speeds persist. NeRF-based pose estimation methods often require a set of images to be processed simultaneously, limiting their applicability in real-time scenarios.

2.2.2 3D Gaussian Splatting

3D Gaussian Splatting[10] is an explicit radiance field technique for efficient and high-quality rendering.[8]

3D GS starts with the sparse SfM point cloud and creates a set of 3D Gaussians, then it optimizes and adaptively controls the density of this set of Gaussians. Once trained, the renderer allows real-time navigation for a wide variety of scenes. In addition, in contrast to NeRF, which relies on computationally expensive volumetric ray sampling, 3DGS achieves real-time rendering through a tile-based rasterizer.[10] For more details on 3DGS and related works, refer to [11], [12].

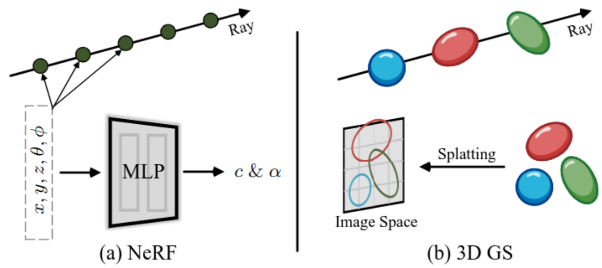

Figure 2: NeRFs vs. 3D GS. Image credits:[11]

As shown in Fig2, NeRF samples along the ray and then queries the MLP to obtain corresponding colors and opacities. In contrast, 3D GS projects all 3D Gaussians into the image space (platting) and then performs parallel rendering, which can be viewed as a forward mapping (splatting and rasterization).[11]

3 Vision for Decision Making

Most conventional robot decision modules or learned policies rely solely on RGB images from multiple cameras as input. However, they exhibit limited or no generalization capabilities, which are essential for real-world robotic applications. To address this limitation, recent research has focused on developing or adapting perception modules specifically designed for robotic tasks. These works generally fall into two main categories: (1) extracting visual features from images to aid downstream robotic tasks, and (2) building dynamic models from visual observations.

3.1 Visual Features for Policy

In computer vision, the term "feature" is broad, but in the context of robot learning, it has three distinct meanings: affordances, features extracted from foundational models, and task-specific visual representations.

Affordance refers to a value map that indicates which parts of an object are most suitable for grasping or are most likely to move under force during actions like pulling or pushing. This concept is extremely useful in the manipulation of articulated objects. Where2Act [13] is a pioneering work that proposes a universal learning-from-interaction framework to derive affordances through online experiences. It bridges visual perception and actionable understanding by predicting interaction points on articulated objects based on pixel data. Building on this, Where2Explore [14] and Robo-ABC [15] extend affordance learning to unseen cases, enabling robots to generalize in a few-shot manner by leveraging local geometric similarities or semantic correspondences between objects. However, most of these works overlook the complexity of occlusions and real-world environments. To address these challenges, Wu et al. [16] focus on enabling robots to manipulate partially visible articulated objects. Finally, the aforementioned works generally assume that the manipulated objects are articulated. To expand affordance learning to deformable objects, Wu et al. [17] propose a novel framework that teaches robots to predict and plan complex manipulations by understanding the visual and physical properties of deformable objects (e.g., ropes and fabrics).

Recently, many foundational models in vision, such as those presented by Radford et al. [18], Oquab et al. [19], Khirodkar et al. [20], and Guzhov et al. [21], have unlocked unprecedented applications for robotics. From a human perspective, most objects within the same category share common local parts and configurations. This similarity facilitates skill transfer in humans. This key insight introduces the concept of generalizable feature fields [22], which ensure that similar objects exhibit similar features [19] when used as observational input. Another key application of foundational models is in robot exploration. To fully leverage CLIP features—which align images with language embeddings—Qiu et al. [23] combine CLIP with SLAM, allowing robots to understand human instructions, explore unknown environments, and perform tasks simultaneously. Jatavallabhula et al. [24], on the other hand, focus on combining features from multiple foundational models, enabling robots to build open-set 3D maps that can be queried using text, clicks, images, or audio.

3.2 Dynamic Model from Vision

Visual perception provides rich information about the target object and its environment, which can be used to learn dynamic models for deformable objects or world models to predict near-future states for planning purposes.

A Graph Neural Network (GNN) is a common choice for modeling deformable objects as a set of particles with internal connections that represent their physical properties. For instance, Shi et al. [25] train a GNN to predict the future shape of dough after applying a chosen tool and a set of randomly sampled actions. This allows the robot to plan sequences of actions and tool combinations to shape the dough into a desired form. GNNs trained from visual observations are not limited to elastic-plastic objects but are also applicable to liquids [26][27] and granular object piles [28].

Rather than focusing on specific objects, a world model is a more general approach that predicts the future state of the environment. In this regard, many works use world models as simulators of real environments, training robot policies within them to reduce the need for time-consuming data collection [29][30][31][32]. However, in these approaches, the video prediction model (powered by the world model) is trained before policy training, and the policy's performance heavily depends on the quality of the world model. An alternative approach is to train the world model and the robot policy simultaneously through reinforcement learning [33][34][35]. This method not only reduces data collection time but also enables online training and exploration [36].

4 Motion Perception

Most robotic tasks operate in 3D space, so in addition to perceiving the environment at the pixel level, understanding the 3D actions of objects and robots is crucial. This understanding can be divided into various levels, from simple to complex, e.g., 1. 6-DoF pose estimation of objects, i.e., perceiving the 6-DoF pose of target objects; 2. Motion re-targeting, i.e., mapping sequential actions to joint space; 3. Video demonstration, i.e., understanding 3D physical interaction videos. We will introduce each task and how each level of understanding helps with robotic tasks

4.1 6-DoF Pose Estimation

6-DoF Pose Estimation [37, 38] refers to estimating the position and orientation of objects simultaneously. The learned pose can be used in a typical downstream robotic task, such as object grasping[39, 40]. With the resulting target 6-DoF pose, robots can compute the desired motion through inverse kinematics, which makes it crucial to robotics. In general, 6-DoF poses can be recovered through invariant features [41, 42, 43] from multiple 2D images taken from different angles, or through template mapping[44] when the CAD models are known. Both approaches have been applied in deep learning frameworks. In the following content under this section, we will focus on those deep learning-based methods. We followed the problem setting categories proposed in [45] where the pose estimation problem is divided into **Instance-level Pose Estimation**, **Category-level Pose Estimation**, and **Unseen Object Pose Estimation**.

4.1.1 Instance-level Pose Estimation

The typical problem setting in instance-level pose estimation is to estimate a specific object, meaning the objects at test time are always as same at training time.

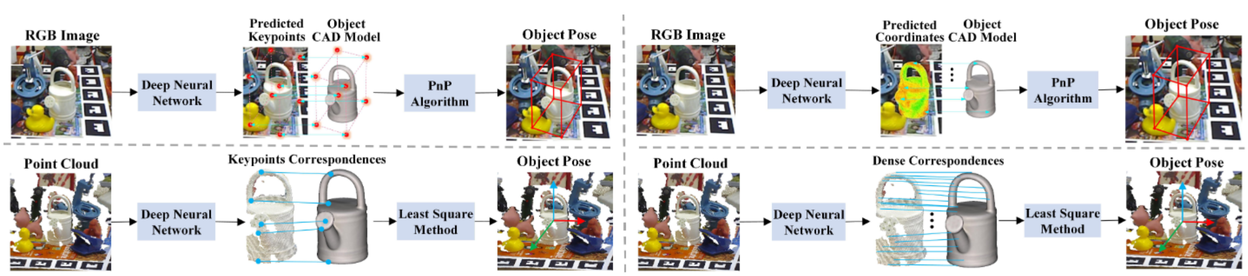

Figure 3: Instance-level correspondence methods pipeline. Sparse correspondence methods (left). Dense correspondence methods (right). Image credits [45]

Most instance-level pose estimation involves known CAD models. Some methods match observed images with known CAD models through correspondence matching[46, 47], where dense correspondence matching always involves matching all pixel coordinates or features, while sparse feature correspondence only matches keypoints[48]. We show sparse and dense correspondence matching in Fig3. When dealing with dense correspondence, recent works often incorporate techniques such as NerF[49] or feature fields [50].

Template methods match by extracting texture-less features, including features from RGB, point cloud, or depth images to attempt to improve robustness. These methods try to find the most similar templates from labeled templates. Based on the feature classes, some works[51, 52] are done on RGB-based features. Others are done in point cloud features[53, 54].

More directly, some methods attempt to perform regression directly on images and target poses. These can be divided by geometry-guided regression and direct regression. Geometry-guided regression often requires more information to digest, such as decoupling 6-DoF poses into translation features and rotation features [55], or learning geometric and contextual features [56]. By contrast, direct regression methods directly output the estimated poses from visual inputs without any other information [57, 58, 59].

4.1.2 Category-level Pose Estimation

Compared to instance-level estimation, category-level estimation can generalize to unseen objects when the objects are within established categories. Category-level methods are mainly developed in two directions, one is shape-prior-based and the other one is shape-prior-free.

For shape-prior-based methods, they either use **Normalized Object Coordinate Space** (NOCS) to align the object point cloud or features to predicted ones, or directly regress on extracted input features. [60] obtained the shape-prior features in offline. Recent work [61] divided the alignment into different parts including coarse deformation, fine deformation, and recurrent refinement. Among them, some methods incorporate self-supervision [62], and semantic features [63] to further increase the robustness and generalization ability.

Shape-prior-free methods[64] do not rely on any priors. Some people [65] incorporate 3D Graph Convolution into the process by leveraging the depth information, which is further improved by [66]. Some introduced auto-encoders [67], or diffusion models [68] into the category-pose estimation problem. [69] first introduced NOCS to generate a canonical representation for objects within one category.

4.1.3 Unseen Object Pose Estimation

Unseen object pose estimation methods can be generalized to unseen objects. There are typically two categories: CAD model-based methods and manual reference view-based methods. Each category can be further divided into feature-matching or template-matching methods.

Typical feature-matching methods in CAD model-based are [70] which involve PnP+RANSAC algorithm[71]; while template-matching methods [72, 73, 74], learn different object poses in latent space. However, these methods require CAD models which hinders their application to wider scenarios.

Recent works [75, 76, 77] relax the assumption on CAD model where only manually pre-captured reference views are required. BundleSDF[78] further relaxes the problem setting by incorporating the process of 3D model reconstruction and 6-DoF pose estimation.

4.2 Human Demonstration

With the success of robotic learning, particularly reinforcement learning and imitation learning, in many tasks, as well as the success of large language models highlighting the importance of data, the question of how to obtain sufficient data for robotic learning is becoming increasingly critical. Since directly teleoperating robots consume both human labor and time, some recent works have sought to learn from human demonstrations. These approaches can be divided into two categories: the first one is based on **human motion re-targeting**, where the action sequences are mapped from the human joint space to the robot's joint space; the second one involves offline learning directly from **videos of human demonstrations**, aiming to gain benefits or even learn action sequences directly from it.

4.2.1 Motion Re-targeting

Motion re-targeting refers to transferring motion from one system (often a human) to another (often a robot or a digital avatar) while preserving the core characteristics of the motion. Previous methods[79, 80, 81] typically map the results obtained from motion capture (MoCap) to the joint space of robots or animated characters. However, due to the

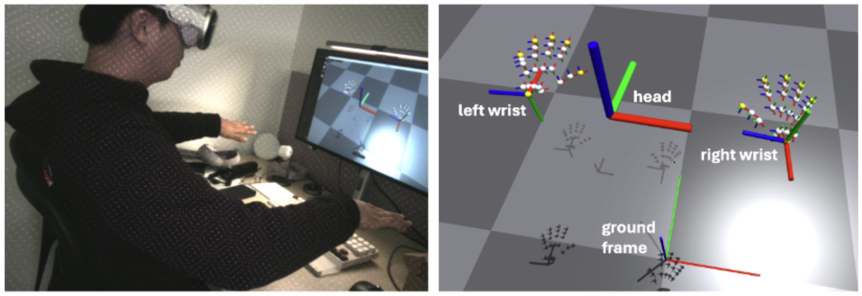

Figure 4: VisionProTeleop System. Users can teleoperate with their robots while seeing the motions in Apple Vision Pro. Image credits: https://github.com/Improbable-AI/VisionProTeleop?tab=readme-ov-file

robots' under-actuation, limited degrees of freedom, and joint constraints, this often involves complex optimization problems. Moreover, the motions generated by joint mapping frequently do not account for their feasibility in real-world environments, making them difficult to apply in robot teleoperation. With the growing attention to Imitation Learning (IL) and the development of physics-based simulation engines, teleoperation methods based on motion re-targeting have gained increasing attention in recent years.

Some full-body control methods [82, 83] map keypoints of the human body to the robot, and then determine if the robot can execute the action. After that, they conduct RL training in a physics engine to enable robots to adapt to physics and finally deploy the strategy through sim2real. Some methods[84, 85] focusing on dexterous manipulation tasks use visual re-targeting entirely. With the help of virtual reality devices, such as Apple Vision Pro, they can achieve re-targeting of various robot finger joints, thus enabling teleoperation.

4.2.2 Video Demonstration

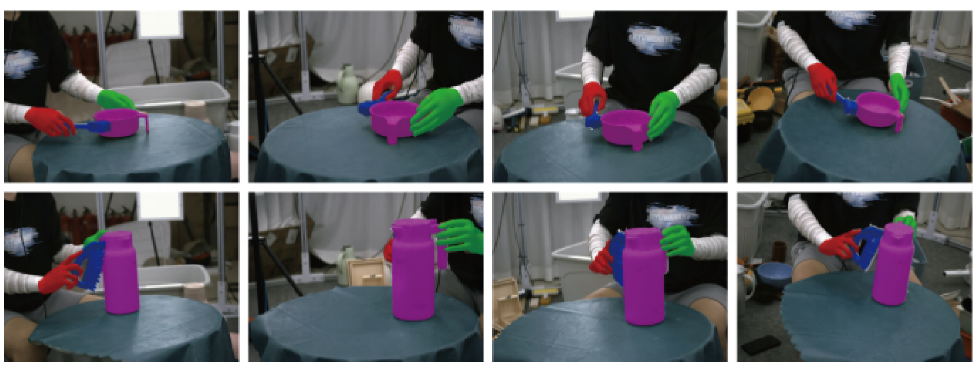

Figure 5: TACO dataset. TACO contains 2.5K motion sequences paired with third-person and egocentric views, precise hand-object 3D meshes, and action labels. Image credits:[86]

The goal of video demonstration is to enable robots to learn action sequences through videos of human demonstrations, instead of through manually collected ground-truth robot states. Some early works[87, 88, 89, 90] attempted to do this through explicit keypoint matching and object pose estimation. Some researchers tried to learn rewards from demonstration[91, 92, 93], i.e., inverse reinforcement learning. Others[94] proposed using a meta-learning framework to solve the problem, aiming to map data points from one domain to another. An interesting work [95]attempted to use generative networks to translate human demonstrations into robot demonstrations directly.

Recently, with the introduction of large datasets of human demonstrations[96, 97, 86], more and more methods have begun to focus on training with large amounts of data in simulators. Here, we show the TACO [86] human demonstration dataset in Fig5. Some works[98, 99, 100] explicitly perform re-targeting between human arms and robot arms for imitation learning. Other works[101, 102, 103] fully use simulated environments for reinforcement learning (RL), as RL can obtain billions of data in simulation environments, thus often learning more expressive and robust actions. The latest works[104] are attempting to use RL and human demonstrations to learn more complex tasks such as bimanual manipulation.

5 Conclusion

In this survey, we have explored the key advancements in 3D vision technologies for robotics, focusing on Perception and Environment Modeling, Motion Perception, and Visual Decision Making. These technologies have enabled robots to achieve higher autonomy by improving their ability to perceive, understand, and interact with complex environments. Recent innovations like NeRF and 3D Gaussian Splatting have significantly enhanced 3D reconstruction capabilities.

Looking ahead, the integration of Large Language Models (LLMs) with 3D vision presents an exciting opportunity to further enhance robot intelligence. By combining LLMs' powerful language understanding capabilities with 3D perception, robots could achieve more intuitive human-robot interaction, process complex instructions, and even collaborate more effectively in dynamic and unstructured environments.

We hope this survey serves as a useful reference for researchers and practitioners looking to further develop robotic systems capable of leveraging both 3D vision and machine intelligence.

References

[1] Basheer Al-Tawil, Thorsten Hempel, Ahmed Abdelrahman, and Ayoub Al-Hamadi. A review of visual slam for robotics: evolution, properties, and future applications. *Frontiers in Robotics and AI*, 11:1347985, 2024.

[2] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE Transactions on Robotics*, page 1147–1163, Sep 2015.

[3] Raul Mur-Artal and Juan D. Tardos. Orb-slam2: an open-source slam system for monocular, stereo and rgb-d cameras. *IEEE Transactions on Robotics*, page 1255–1262, Sep 2017.

[4] Carlos Campos, Richard Elvira, Juan J. Gomez Rodriguez, Jose M. M. Montiel, and Juan D. Tardos. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, page 1874–1890, Nov 2021.

[5] Anton Kasyanov, Francis Engelmann, Jorg Stuckler, and Bastian Leibe. Keyframe-based visual-inertial online slam with relocalization. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Aug 2017.

[6] Raul Mur-Artal and Juan D. Tardos. Visual-inertial monocular slam with map reuse. *IEEE Robotics and Automation Letters*, page 796–803, Mar 2017.

[7] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, Oct 2011.

[8] Youmin Zhang Fabio Tosi, Ziren Gong, Stefano Mattoccia Erik Sandström, Martin R. Oswald, and Matteo Poggi. How nerfs and 3d gaussian splatting are reshaping slam: a survey. *arXiv preprint arXiv:2402.13255v2*, 2024.

[9] 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. In *ECCV*, 2020.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[11] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.

[12] Tong Wu, Yu-Jie Yuan, Ling-Xiao Zhang, Jie Yang, Yan-Pei Cao, Ling-Qi Yan, and Lin Gao. Recent advances in 3d gaussian splatting. *Computational Visual Media*, 10(4):613–642, 2024.

[13] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[14] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. *arXiv preprint arXiv:2401.07487*, 2024.

[16] Ruihai Wu, Kai Cheng, Yan Zhao, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10947–10956, 2023.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[20] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models, 2024.

[21] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021.

[22] Yixuan Wang, Guang Yin, Binghao Huang, Tarik Kelestemur, Jiuguang Wang, and Yunzhu Li. GenDP: 3d semantic fields for category-level generalizable diffusion policy. In *8th Annual Conference on Robot Learning*, 2024.

[23] Ri-Zhao Qiu, Yafei Hu, Ge Yang, Yuchen Song, Yang Fu, Jianglong Ye, Jiteng Mu, Ruihan Yang, Nikolay Atanasov, Sebastian Scherer, and Xiaolong Wang. Learning generalizable feature fields for mobile manipulation, 2024.

[24] Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

[25] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools, 2023.

[26] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids, 2019.

[27] Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields, 2022.

[28] Yixuan Wang, Yunzhu Li, Katherine Driggs-Campbell, Li Fei-Fei, and Jiajun Wu. Dynamic-resolution model learning for object pile manipulation, 2023.

[29] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination, 2024.

[30] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Dee M, Jodilyn Peralta, Brian Ichter, Karol Hausman, and Fei Xia. Scaling robot learning with semantically imagined experience, 2023.

[31] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023.

[32] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences, 2023.

[33] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024.

[34] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination, 2020.

[35] Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models, 2021.

[36] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer: World models for physical robot learning, 2022.

[37] Alvaro Collet and Siddhartha S Srinivasa. Efficient multi-view object recognition and full pose estimation. In *2010 IEEE International Conference on Robotics and Automation*, Apr 2010.

[38] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. *The International Journal of Robotics Research*, page 1284–1306, Aug 2011.

[39] Boshi An, Yiran Geng, Kai Chen, Xiaoqi Li, Qi Dou, and Hao Dong. Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7748–7755. IEEE, 2024.

[40] Snehal Jauhri, Sophie Lueth, and Georgia Chalvatzaki. Active-perceptive motion generation for mobile manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1413–1419. IEEE, 2024.

[41] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004.

[42] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. *SURF: Speeded Up Robust Features*, page 404–417. Dec 2005.

[43] Steven S. Beauchemin and John L. Barron. The computation of optical flow. *ACM computing surveys (CSUR)*, 27(3):433–466, 1995.

[44] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *2011 International Conference on Computer Vision*, Oct 2011.

[45] Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: A comprehensive survey, 2024.

[46] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017.

[47] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018.

[48] Shuxuan Guo, Yinlin Hu, Jose M Alvarez, and Mathieu Salzmann. Knowledge distillation for 6d pose estimation by aligning distributions of local predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18633–18642, 2023.

[49] Fu Li, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2123–2133, 2023.

[50] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: 6-dof object pose estimation via recurrent correspondence field estimation and pose optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[51] Zhigang Li and Xiangyang Ji. Pose-guided auto-encoder and feature-based refinement for 6-dof object pose regression. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8397–8403. IEEE, 2020.

[52] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021.

[53] Haobo Jiang, Mathieu Salzmann, Zheng Dang, Jin Xie, and Jian Yang. Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[54] Zheng Dang, Lizhou Wang, Yu Guo, and Mathieu Salzmann. Learning-based point cloud registration for 6d object pose estimation in the real world. In *European conference on computer vision*, pages 19–37. Springer, 2022.

[55] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. 6d object pose regression via supervised learning on point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3643–3649. IEEE, 2020.

[56] Yifei Shi, Junwen Huang, Xin Xu, Yifan Zhang, and Kai Xu. Stablepose: Learning 6d object poses from geometrically stable patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15222–15231, 2021.

[57] Xiaoke Jiang, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, and Liwei Wu. Uni6d: A unified cnn framework without projection breakdown for 6d pose estimation. ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 5, 2022.

[58] Yang Hai, Rui Song, Jiaojiao Li, and Yinlin Hu. Shape-constraint recurrent flow for 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4831–4840, 2023.

[59] Yuelong Li, Yafei Mao, Raja Bala, and Sunil Hadap. Mrc-net: 6-dof pose estimation with multiscale residual correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10476–10486, 2024.

[60] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020.

[61] Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. Catformer: Category-level 6d object pose estimation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6808–6816, 2024.

[62] Yisheng He, Haoqiang Fan, Haibin Huang, Qifeng Chen, and Jian Sun. Towards self-supervised category-level object pose and size estimation. *arXiv preprint arXiv:2203.02884*, 2022.

[63] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021.

[64] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021.

[65] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021.

[66] Jierui Liu, Zhiqiang Cao, Yingbo Tang, Xilong Liu, and Min Tan. Category-level 6d object pose estimation with structure encoder and reasoning attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6728–6740, 2022.

[67] Xinke Deng, Junyi Geng, Timothy Bretl, Yu Xiang, and Dieter Fox. icaps: Iterative category-level object pose and shape estimation. *IEEE Robotics and Automation Letters*, 7(2):1784–1791, 2022.

[68] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[69] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[70] Giorgia Pitteri, Slobodan Ilic, and Vincent Lepetit. Cornet: generic 3d corners for 6d pose estimation of new objects without retraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[71] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[72] Vassileios Balntas, Andreas Doumanoglou, Caner Sahin, Juil Sock, Rigas Kouskouridas, and Tae-Kyun Kim. Pose guided rgbd feature learning for 3d object pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 3856–3864, 2017.

[73] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O Arras, and Rudolph Triebel. Multi-path learning for object pose estimation across domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13916–13925, 2020.

[74] Yilin Wen, Xiangyu Li, Hao Pan, Lei Yang, Zheng Wang, Taku Komura, and Wenping Wang. Disp6d: Disentangled implicit shape and pose learning for scalable 6d pose estimation. In *European Conference on Computer Vision*, pages 404–421. Springer, 2022.

[75] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, Wenping Wang, and Hong Kong. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images.

[76] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2020.

[77] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. May 2022.

[78] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023.

[79] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

[80] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

[81] Xue Bin Peng, Glen Berseth, KangKang Yin, and Michiel Van De Panne. Deeploco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Acm transactions on graphics (tog)*, 36(4):1–13, 2017.

[82] Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. *arXiv preprint arXiv:2403.04436*, 2024.

[83] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[84] Younghyo Park and Pulkit Agrawal. Using apple vision pro to train and control robots, 2024.

[85] Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning, 2024.

[86] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding, 2024.

[87] Yasuo Kuniyoshi, Masayuki Inaba, and Hirochika Inoue. Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE transactions on robotics and automation*, 10(6):799–822, 1994.

[88] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 61(12):1323–1334, 2013.

[89] Karinne Ramirez-Amaro, Michael Beetz, and Gordon Cheng. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artificial Intelligence*, 247:95–118, 2017.

[90] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[91] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016.

[92] Bradly C Stadie, Pieter Abbeel, and Ilya Sutskever. Third-person imitation learning. *arXiv preprint arXiv:1703.01703*, 2017.

[93] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1118–1125. IEEE, 2018.

[94] Tianhe Yu, Chelsea Finn, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

[95] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

[96] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Learning-from-demonstrations benchmark for generalizable manipulation skills. *CoRR, abs/2107.14483, 2021b. URL https://arxiv. org/abs/2107*, 14483, 2021.

[97] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596*, 2023.

[98] SridharPandian Arunachalam, Sneha Silwal, Ben Evans, and Lerrel Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation.

[99] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023.

[100] Qingtao Liu, Yu Cui, Qi Ye, Zhengnan Sun, Haoming Li, Gaofeng Li, Lin Shao, and Jiming Chen. Dexrepnet: Learning dexterous robotic grasping network with geometric and spatial hand-object representations. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3153–3160. IEEE, 2023.

[101] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. Vividex: Learning vision-based dexterous manipulation from human videos. *arXiv preprint arXiv:2404.15709*, 2024.

[102] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021.

[103] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research*, 40(12-14):1419–1434, 2021.

[104] Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations, 2024.