자연어 기반 로봇 파지 연구 동향 및 제안

1. 서론

1.1 연구 배경 및 필요성

오늘날 자율 로봇 시스템은 다양한 분야에서 그 중요성이 점차 증대되고 있으며, 이러한 로봇이 실제 환경과 효과적으로 상호작용하기 위해서는 물체를 안전하고 정확하게 파지하는 능력이 필수적이다. 특히, 인간과 로봇 간의 협업이 증가함에 따라, 자연어 인터페이스를 통해 로봇을 직관적으로 제어하고 원하는 작업을 수행하도록 하는 연구의 필요성이 강조되고 있다. 본 보고서는 자연어 명령을 기반으로 로봇이 물체를 파지하는 두 가지 주요 연구 방향, 즉 unlabeled 객체 파지와 비전-언어-행동 (Vision-Language-Action, VLA) 모델 기반 파지에 대한 최신 동향을 분석하고, 향후 연구 방향을 제시하고자 한다. Unlabeled 객체 파지는 로봇이 사전에 특정 객체에 대한 정보를 가지고 있지 않은 상태에서 자연어 명령을 이해하고 해당 속성을 가진 객체를 인식하여 파지하는 기술을 의미하며, VLA 기반 파지는 시각적 정보와 자연어 명령을 통합적으로 처리하여 로봇의 파지 행동을 직접 생성하는 end-to-end 학습 방식에 기반한다. 이러한 연구들은 서비스 로봇, 제조 자동화, 재난 구조 등 다양한 분야에서 로봇의 활용 가능성을 크게 확장할 수 있을 것으로 기대된다. 본 보고서에서는 이러한 연구를 위해 사용될 주요 하드웨어인 RGBD 카메라와 Jetson AGX Orin 64GB 모델의 특징과 장점을 간략히 소개하고, 각 연구 방향에 대한 심층적인 분석을 제공할 것이다.

1.2 본 보고서의 목표 및 범위

본 보고서는 자연어 기반 로봇 파지 연구 분야에서 연구 주제를 선정하고자 하는 사용자의 요구에 부응하기 위해 다음과 같은 목표를 설정하고 범위를 한정한다. 첫째, unlabeled 객체 인식 및 파지를 위한 최신 알고리즘 및 기술 동향을 분석하고, 특히 사용자가 관심을 가지고 있는 CLIP 알고리즘 외에 RGBD 카메라 입력에 적합한 다른 대안들을 비교한다. 둘째, 자연어 명령 기반 로봇 파지 (VLA) 모델의 작동 방식과 특징을 비교 분석하여 사용자의 이해를 돕는다. 셋째, RGBD 데이터를 이용하여 로봇 팔이 파지하기 적절한 지점을 결정하는 다양한 방법들을 조사하고, 각 방법의 정확성, 효율성, 그리고 MoveIt 2와의 통합 가능성을 평가한다. 넷째, 인식된 객체의 파지 지점 정보를 MoveIt 2에 효과적으로 전달하여 모션 계획 및 실행을 가능하게 하는 인터페이스 및 데이터 형식에 대한 정보를 검색한다. 다섯째, Jetson AGX Orin 환경에서 실시간으로 객체 인식 및 파지 작업을 수행하기 위한 최적화 기법들을 조사하고, 각 알고리즘 및 기술의 연산 요구 사항과 실제 하드웨어에서의 성능을 예측할 수 있는 자료를 분석한다. 여섯째, Isaac Sim 환경에서 강화 학습을 사용하여 로봇 팔의 파지 동작을 학습시키는 일반적인 방법들을 조사하고, SAC 및 PPO 알고리즘을 continuous action space에 적용하는 구체적인 사례 및 고려 사항들을 제시한다. 마지막으로, 첫 번째 연구 주제 (unlabeled 객체 인식 및 파지)와 두 번째 연구 주제 (VLA 기반 파지)의 핵심적인 차이점과 유사점을 분석하고, 각 접근 방식의 장단점을 비교하여 사용자의 연구 목표에

더 적합한 방향을 제시한다.

1.3 보고서 구성

본 보고서는 서론에 이어, 2장에서는 자연어 기반 unlabeled 객체 파지 연구의 최신 알고리즘 동향, CLIP과 로봇 팔 파지 작업의 통합 방안, RGBD 데이터를 이용한 파지 지점 결정 방법, MoveIt 2를 활용한 모션 계획 및 실행, 그리고 Jetson AGX Orin 환경에서의 실시간 성능 검증 및 고려 사항을 상세히 다룬다. 3장에서는 자연어 명령 기반 로봇 파지 연구인 VLA 모델의 최신 연구 동향과 다양한 VLA 모델의 작동 방식 및 특징을 비교 분석한다. 4장에서는 앞서 논의된 두 가지 연구 주제의 핵심적인 차이점과 유사점을 분석하고, 각 접근 방식의 장단점을 비교하여 사용자의 연구 목표에 따른 최적 연구 방향을 제시한다. 5장에서는 Isaac Sim 환경에서 강화 학습을 사용하여 로봇 팔의 파지 동작을 학습시키는 일반적인 방법과 SAC 및 PPO 알고리즘의 적용 전략을 설명한다. 마지막으로 6장에서는 본 보고서의 주요 연구 결과를 요약하고 향후 연구를 위한 제언을 제시하며 보고서를 마무리한다.

2. 자연어 기반 **Unlabeled** 객체 파지 연구

2.1 Unlabeled 객체 인식을 위한 최신 알고리즘 동향

2.1.1 CLIP 알고리즘 분석 및 로봇 파지 적용 가능성

CLIP (Contrastive Language-Image Pre-training) 모델은 이미지와 텍스트 간의 의미적 유사성을 학습하는 Vision-Language Model (VLM)로, 대규모 이미지-텍스트 쌍 데이터셋으로 사전 학습되어 풍부한 시각적 및 언어적 이해 능력을 갖추고 있다.[1] CLIP의 핵심 특징 중 하나는 zero-shot 학습 능력인데, 이는 학습 데이터에 존재하지 않았던 새로운 객체에 대해서도 텍스트 프롬프트를 이용하여 이미지의 내용을 분류할 수 있다는 점이다.[1] 이러한 능력은 로봇이 사전에 특정 객체에 대한 명시적인 학습 없이도 자연어 설명을 통해 이해하고 작업을 수행할 수 있는 잠재력을 시사한다. 실제로 로봇 파지 작업에 CLIP을 활용하는 연구에서는 자연어 명령 (예: "빨간색 컵")을 기반으로 특정 속성을 가진 객체를 인식하는 가능성을 탐색하고 있다.[45]에서는 속성 기반 로봇 파지 접근 방식이 제시되며, RGB-D 카메라를 통해 얻은 시각 정보와 텍스트 쿼리를 융합하여 파지 가능성을 예측하는 모델을 제안한다.

CLIP의 zero-shot 능력은 unlabeled 객체 파지 분야에서 매우 유망한 접근 방식이지만, 실제 로봇 환경에서의 viewpoint 변화나 occlusion과 같은 문제에 대한 robustness는 추가적인 연구를 통해 확보해야 할 부분이다.[33]에서는 CLIP이 다양한 viewpoint와 occlusion 수준에 민감하게 반응할 수 있음을 지적하며, active recognition과 같은 방법을 통해 이러한 한계를 극복할 필요성을 제기한다. Active recognition은 로봇이 능동적으로 다양한 시점에서 객체를 관찰하고 정보를 수집하여 인식 성능을 향상시키는 전략이다.

2.1.2 RGBD 카메라 입력에 적합한 **CLIP** 외 대안 알고리즘 비교

CLIP 외에도 RGBD 카메라 입력을 활용하여 unlabeled 객체를 인식하고 파지 작업에 적용할 수 있는 다양한 대안 알고리즘들이 존재한다. 이러한 알고리즘들은 각기 다른 특징과 장단점을 가지고 있으며, 특정 연구 목표나 환경에 따라 CLIP보다 더 적합할 수 있다.

개방형 어휘 객체 검출 모델 (Open-Vocabulary Object Detection Models)은 학습 시에 보지 못했던 새로운 객체 클래스를 인식할 수 있도록 설계된 모델들이다. YOLO-World는 이러한 개방형 어휘 객체 검출 모델 중 하나로, 다양한 객체 클래스를 인식하는 데 활용될 수 있다.[6][6]에서는 YOLO-World가 CLIP과 함께 개방형 어휘 분류 및 객체 검출에 사용될 수 있음을 언급하며, 이는 로봇이 다양한 종류의 unlabeled 객체를 인식하는 데 유용할 수 있음을 시사한다. OWL-ViT 역시 텍스트 쿼리를 기반으로 사전 학습 없이 객체 검출을 수행하는 모델로 [1], 자연어 명령에 따라 특정 객체를 찾는 로봇 작업에 적용될 수 있다.

RGBD 데이터를 직접적으로 활용하여 3D 장면에서 새로운 객체를 발견하고 분류하기 위한 프레임워크인 CoDA (Collaborative Novel Box Discovery and Cross-modal Alignment)는 2D 시맨틱 정보와 3D geometry 정보를 융합하여 pseudo label을 생성하고, cross-modal alignment를 통해 특징 공간을 정렬하는 방식을 사용한다.[7][8]에서는 CoDA가 제한된 기본 카테고리 정보만으로 3D 장면에서 새로운 객체를 локализация하고 분류하는 것을 목표로 하며, 2D 시맨틱 정보와 3D geometry 정보를 활용하여 pseudo label을 생성하고, cross-modal alignment를 통해 특징 공간을 정렬하는 방식을 제시한다. [8]은 CoDA의 핵심 방법론과 실험 결과를 요약 제공하며, SUN-RGBD 및 ScanNet 데이터셋에서 기존 방법 대비 상당한 성능 향상을 보임을 강조한다. CoDA는 CLIP과 유사하게 텍스트와 이미지를 활용하지만, 3D 공간에서의 객체 인식에 특화되어 있어 로봇 파지에 더 적합할 수 있으며, 2D와 3D 공간의 특징을 융합하여 새로운 객체를 효과적으로 локализация하고 분류하는 능력은 중요한 장점이다.[8][8]에서도 CoDA의 접근 방식이 2D와 3D 공간의 cross-prior를 활용하여 새로운 3D 객체를 발견하고 localization하는 효과적인 전략임을 긍정적으로 평가한다.

OpenNav는 스마트 휠체어 네비게이션을 위해 개발된 효율적인 개방형 어휘 3D 객체 검출 파이프라인으로, RGB-D 이미지를 기반으로 작동한다.[12][14]에서는 OpenNav가 RGB-D 이미지를 입력으로 받아 개방형 어휘 2D 객체 검출기, 마스크 생성기, 깊이 분리, point cloud 재구성을 통합하여 3D bounding box를 생성하는 파이프라인을 설명한다. [13]는 OpenNav의 핵심 아이디어를 평이하게 설명하며, 언어 모델을 활용하여 사용자가 자연어로 묘사하는 다양한 객체를 검출할 수 있음을 강조한다. [12]는 OpenNav 관련 정보를 제공하는 웹사이트가 접근 불가능함을 나타낸다. [16]은 OpenNav의 공식 코드를 제공하는 GitHub 저장소를 소개한다. OpenNav는 실시간성과 효율성에 초점을 맞추고 있어, Jetson AGX Orin과 같은 임베디드 환경에서 실행 가능성을 보여주며, 2D 객체 검출과 깊이 정보를 결합하여 3D bounding box를 생성하는 방식은 파지 작업에 유용할 수 있다.[14][17]에서는 OpenNav가 Replica 데이터셋에서 state-of-the-art 성능을 달성했음을

보여준다.

Zero-Shot Object Detection 모델은 학습 데이터 없이 새로운 객체 클래스를 인식하는 것을 목표로 한다. GroundingDINO는 텍스트 설명을 기반으로 임의의 객체를 검출할 수 있는 Transformer 기반 모델로 [18], 자연어 명령에 따라 특정 객체를 파지하는 로봇 작업에 활용될 수 있다. Segment Anything Model (SAM)은 대규모 이미지 데이터셋으로 학습된 고품질 객체 마스크 생성 모델로, zero-shot 학습 능력을 갖추고 있어 [18], CLIP과 함께 사용되어 객체 인식 및 분할 성능을 향상시키고, 특히 알려지지 않은 객체에 대한 처리 능력을 강화할 수 있다. [18]에서는 GroundingDINO가 언어와 시각 정보를 융합하여 개방형 어휘 개념 일반화를 가능하게 하는 모델임을 설명하고, SAM이 transformer 기반 모델이며, 학습 중에 보지 못한 이미지에 대해서도 객체 마스크를 생성할 수 있는 zero-shot 학습 능력을 갖추고 있음을 설명한다.

자연어 속성 (색상, 모양, 카테고리 이름 등)을 이용하여 객체를 인식하고 파지하는 **속성 기반 파지 모델 (Attribute-Based Grasping Models)**은 새로운 객체에 대한 일반화 능력이 뛰어나며, 사용자 명령의 다양성을 처리하는 데 유리하다.[45]에서는 속성 기반 로봇 파지 접근 방식이 새로운 객체에 대한 일반화 및 적응 능력을 향상시킬 수 있음을 강조하며, RGB-D 카메라를 이용하여 사용자의 명령에 따라 특정 속성을 가진 객체를 파지하는 시스템을 제시한다. [20]은 해당 연구의 저자 및 관련 논문 정보를 제공하며[23]은 해당 논문의 arXiv 정보를 제공한다. [22]은 해당 논문의 제목 일부를 보여준다. [4]은 해당 웹사이트가 접근 불가능함을 나타낸다. [4]에서 제시된 연구는 RGB-D 카메라를 이용하여 속성 기반 파지를 구현하고, 데이터 효율적인 적응 능력을 보여주며, 특히 적대적 적응 및 one-grasp 적응 방법을 통해 적은 데이터로도 성능 향상을 이룰 수 있음을 보여준다.

CLIP 외에도 SigLIP, MetaCLIP, AltCLIP, RemoteCLIP, BioCLIP, MobileCLIP, BLIP, BLIPv2, ALBEF, FastViT 등 다양한 이미지 분류 및 임베딩 모델들이 존재하며 [24], 각 모델은 특정 task나 데이터셋에 더 적합할 수 있다. 예를 들어, MobileCLIP은 모바일 환경에서의 효율성을 강조하며, 사용자의 연구 목표와 하드웨어 환경을 고려하여 CLIP 외의 다른 VLM을 탐색하는 것도 좋은 선택지가 될 수 있다.

표 **1: Unlabeled** 객체 인식을 위한 **CLIP** 외 대안 알고리즘 비교

알고리즘 이름	핵심 기술	입력 방식	주요 특징	로봇 파지에 대한 잠재적 장점	잠재적 단점/제한 사항
YOLO-World	개방형 어휘 객체 검출	RGB	다양한 객체 클래스 인식 가능	다양한 unlabeled 객체 인식에 활용 가능	정확도 측면에서 다른 모델 대비**劣位**

					가능성
OWL-ViT	Zero-Shot 객체 검출	RGB	텍스트 쿼리 기반 사전 학습 없이 객체 검출	특정 객체에 대한 명시적 학습 없이 자연어 명령 기반 검출 가능	RGB 입력에 의존적이며, 깊이 정보 활용 어려움
CoDA	개방형 어휘 3D 객체 검출 및 분류	RGBD	2D 시맨틱 정보와 3D geometry 융합, 새로운 객체 효과적 локализация 및 분류	3D 공간에서의 객체 인식에 특화, 로봇 파지에 직접적 활용 가능성 높음	복잡한 구조, 실시간 성능 확보를 위한 최적화 필요
OpenNav	효율적인 개방형 어휘 3D 객체 검출	RGBD	실시간성 및 효율성 강조, 2D 검출과 깊이 정보 결합 3D bounding box 생성	Jetson AGX Orin과 같은 임베디드 환경에서 실시간 실행 가능성, 파지 작업에 유용한 3D 정보 제공	특정 작업 (휠체어 네비게이션)에 최적화되었을 가능성
GroundingDINO	Zero-Shot 객체 검출	RGB	텍스트 설명을 기반으로 임의의 객체 검출	자연어 명령 기반 다양한 객체 검출 가능	RGB 입력에 의존적이며, 깊이 정보 활용 어려움
SAM	Zero-Shot 객체 마스크 생성	RGB	대규모 데이터 학습, zero-shot으로 다양한 객체 마스크 생성	CLIP 등과 결합하여 객체 인식 및 분할 성능 향상, 알려지지 않은 객체 처리 능력 강화	RGB 입력에 의존적이며, 깊이 정보 활용 어려움
속성 기반	속성 기반	RGBD	자연어 속성	새로운	속성 정보

파지 모델	객체 인식 및 파지		(색상, 모양 등) 이용하여 객체 인식 및 파지	객체에 대한 뛰어난 일반화 능력, 사용자 명령 다양성 처리 유리	추출 및 파지 동작 연결에 대한 추가 연구 필요
CLIP Alternatives	이미지 분류 및 임베딩	RGB	다양한 모델 존재 (SigLIP, MetaCLIP 등), 특정 task나 데이터셋에 더 적합 가능	CLIP 외의 다양한 VLM 옵션 제공, 특정 환경에 더 적합한 모델 선택 가능성	CLIP 대비 성능 우위는 task에 따라 다를 수 있음

2.2 CLIP과 로봇 팔 파지 작업의 효과적인 통합 방안

CLIP은 강력한 zero-shot 능력을 바탕으로 로봇 팔 파지 작업의 성능을 향상시키는 데 다양하게 활용될 수 있다. 자연어 명령을 기반으로 객체를 인식하고, 파지 가능성을 예측하며, 심지어 암시적인 명령에 따라 파지 대상을 추론하는 연구들이 활발히 진행되고 있다.

CLIP의 의미적 이해 능력을 로봇 제어에 직접적으로 통합한 CLIPort는 CLIP의 광범위한 의미적 이해 능력과 TransporterNet의 공간적 정밀도를 결합하여 자연어 명령 기반의 다양한 테이블탑 작업을 수행할 수 있는 end-to-end 프레임워크이다.[26] CLIPort는 별도의 객체 포즈 추정이나 분할 없이도 파지 작업을 수행할 수 있음을 보여주며, 이는 복잡한 전처리 과정 없이도 자연어 명령을 로봇 행동으로 직접 연결할 수 있는 가능성을 제시한다.

언어 기반 파지 검출 연구에서는 CLIP을 이용하여 자연어 명령에 따라 객체의 파지 포즈를 검출하는 방법을 탐구한다.[27] Grasp-Anything++라는 대규모 언어 기반 파지 데이터셋은 CLIP과 같은 모델을 학습시켜 자연어 명령에 따른 파지 작업을 가능하게 한다.[28] 이 데이터셋은 파트 수준 및 객체 수준의 상세한 파지 instruction을 제공하여 로봇이 더 정확하고 상황에 맞는 파지 동작을 수행할 수 있도록 지원한다. 또한, 로봇 행동 데이터로 CLIP을 fine-tuning한 Robotic-CLIP 모델은 언어 기반 파지 검출 등 다양한 로봇 작업에서 성능 향상을 보인다.[27]

Multimodal LLM과 CLIP을 통합하여 암시적인 자연어 명령에 따라 객체를 파지하는 reasoning grasping 연구도 주목할 만하다.[29] 이러한 연구들은 CLIP의 언어 이해 능력을 활용하여 단순한 객체 인식뿐만 아니라 상황에 따른 추론 기반 파지 작업을 가능하게 한다. 예를 들어, "마실 물이 필요해"라는 명령에 대해 "컵"이라는 단어가 명시적으로

언급되지 않았음에도 불구하고 로봇이 컵을 인식하고 파지할 수 있도록 하는 것이다.

뿐만 아니라, CLIP의 시각적 특징 추출 능력을 활용하여 다양한 객체에 대한 파지 품질을 예측하고, 이를 통해 더 안정적인 파지 전략을 수립하는 연구도 진행되고 있다.[33] 파지 품질 예측은 로봇이 여러 가능한 파지 자세 중에서 가장 성공률이 높은 자세를 선택하는 데 도움이 된다.

2.3 RGBD 카메라 데이터를 이용한 로봇 팔 파지 지점 결정 방법

RGBD 카메라 데이터를 이용하여 로봇 팔이 파지하기 적절한 지점을 결정하는 것은 성공적인 파지 작업의 핵심 요소이다. 다양한 방법들이 개발되어 왔으며, 각 방법은 정확성, 효율성, 그리고 MoveIt 2와의 통합 가능성 측면에서 서로 다른 특징을 가진다.

Generative Grasping Convolutional Neural Networks (GG-CNN)은 RGBD 이미지에서 각 픽셀에 대한 파지 품질, 각도, 그리퍼 너비를 예측하는 실시간 객체 독립적 파지 합성 방법이다.[33] GG-CNN은 빠른 속도로 파지 지점을 예측할 수 있어 실시간 제어에 적합하며, 예측된 파지 파라미터 (위치, 각도, 너비)를 MoveIt 2의 Grasp 메시지 형식으로 변환하여 파지 동작 계획을 수립할 수 있다.

Grasping Rectangle 표현 방식은 RGBD 이미지에서 객체의 위치, 방향, 그리퍼 개방 너비를 고려한 파지 사각형을 학습하는 방법이다.[35] 이 방법은 다양한 형태의 객체에 대한 파지 표현에 효과적이며, 2단계 학습 프로세스를 통해 효율성과 정확성을 높인다. MoveIt 2에서 파지 사각형 정보를 이용하여 파지 동작 계획을 수립할 수 있으며, 예를 들어, 파지 사각형의 중심점을 파지 목표 위치로 설정하고, 방향 정보를 이용하여 그리퍼의 접근 방향을 결정할 수 있다.

Deep Learning 기반 파지 검출 네트워크는 RGB 및 깊이 정보를 융합하여 파지 가능 영역을 예측하는 U-Net 구조의 CNN과 같은 모델들을 포함한다.[37] 이러한 딥러닝 기반 방법들은 복잡한 환경에서의 파지 작업에 강력한 성능을 보이며, RGBD 데이터의 다양한 정보를 효과적으로 활용할 수 있다. MoveIt 2는 point cloud 데이터를 입력으로 받아 파지 계획을 수립할 수 있으므로, 이러한 네트워크의 출력을 변환하여 연동할 수 있다. 예를 들어, 예측된 파지 가능 영역에서 가장 높은 확률 값을 갖는 픽셀의 3D 좌표를 파지 목표 위치로 설정할 수 있다.

Affordance 기반 파지 검출은 객체의 affordance (어떤 행동을 가능하게 하는 속성)를 인식하여 파지 지점을 결정하는 방법이다.[40] 이 방법은 객체의 기능적 측면을 고려하여 파지 지점을 결정하므로, 단순한 형태 기반 파지보다 더 안정적이고 효율적인 파지를 가능하게 한다. CLIP과 같은 VLM을 이용하여 affordance를 인식하고, MoveIt 2에 적절한 파지 목표를 설정할 수 있으며, 예를 들어, "컵의 손잡이를 잡으라"는 명령에 대해 컵의 손잡이 영역을 affordance로 인식하고, 해당 영역을 파지 목표로 설정할 수 있다.

2.4 MoveIt 2를 활용한 모션 계획 및 실행

MoveIt 2는 로봇 팔의 모션 계획 및 실행을 위한 강력한 프레임워크로, 인식된 객체의 파지 지점 정보를 효과적으로 전달하고 이를 기반으로 안전하고 효율적인 모션 계획을 수립하는 데 필수적인 도구이다.

파지 지점 정보를 MoveIt 2에 전달하는 방법 중 하나는 MoveIt에서 파지 작업을 정의하는 표준 메시지 형식인 moveit_msgs::Grasp 메시지를 활용하는 것이다.[47] 이 메시지는 파지 전/후의 그리퍼 자세, 파지 포즈, 접근/후퇴 방향 및 거리 등을 포함하며, 인식된 객체의 파지 지점 (위치 및 방향) 정보를 grasp_pose 필드에 담고, 그리퍼의 열림/닫힘 상태를 pre_grasp_posture 및 grasp_posture 필드에 정의하여 MoveIt 2에 전달할 수 있다. 또한, 접근 및 후퇴 궤적을 정의하여 안전하고 효율적인 파지 동작 계획을 수립할 수 있다.

인식된 객체를 MoveIt 2의 Planning Scene에 Collision Object로 추가하는 것도 중요한 단계이다.[50] 객체의 3D 모델 또는 bounding box 정보를 이용하여 Planning Scene을 업데이트하고, 파지 동작 계획 시 충돌을 방지할 수 있으며, 특히 복잡한 환경에서 안전한 파지 작업을 위해 필수적이다.

외부에서 개발된 파지 계획 알고리즘 (예: GG-CNN)의 결과를 MoveIt 2에서 사용할 수 있도록 Grasp Generator Plugin을 개발하여 통합하는 방법도 있다.[52] 이를 통해 개발자는 자신만의 파지 알고리즘을 MoveIt 2의 모션 계획 파이프라인에 seamless하게 통합하여 활용할 수 있으며, 다양한 Grasp Generator Plugin들이 존재하므로, 사용자는 자신의 파지 알고리즘에 맞춰 적절한 플러그인을 선택하거나 직접 개발할 수 있다.

MoveIt의 MoveGroup 인터페이스를 통해 제공되는 pick() 액션을 사용하여 파지 작업을 수행할 수도 있다.[47] moveit_msgs::Grasp 메시지 벡터를 인자로 전달하여 다양한 파지 시도를 할 수 있으며, pick() 액션은 파지 전 접근, 파지, 들어올리기 등의 단계를 자동으로 처리하므로, 사용자는 파지 목표만 설정하면 된다. 이는 복잡한 파지 동작을 간단하게 구현할 수 있도록 지원한다.

MoveIt 2의 Perception Pipeline을 활용하여 외부 센서 (RGBD 카메라)로부터 얻은 point cloud 데이터를 처리하고, 인식된 객체 정보를 Planning Scene에 반영하는 것도 가능하다.[51] 이를 통해 실시간으로 변화하는 환경에 대한 인식 결과를 바탕으로 파지 계획을 업데이트하고 실행할 수 있으며, Perception Pipeline은 로봇이 동적인 환경에서 안전하고 효과적으로 작동하는 데 필수적인 기능이다.

2.5 Jetson AGX Orin 환경에서의 실시간 성능 검증 및 고려 사항

Jetson AGX Orin은 최대 275 TOPS의 고성능 AI 컴퓨팅 능력과 15W ~ 60W의 저전력 소비를 특징으로 하는 임베디드 시스템으로 [65], 로봇 파지 작업에 필요한 연산 능력을 제공한다. Jetson AGX Orin 환경에서 실시간 객체 인식 및 파지 작업을 수행하기 위해서는 모델 경량화 (MobileNet, EfficientNet 등) [68], NVIDIA의 딥러닝 추론 최적화

라이브러리인 TensorRT 활용 [66], GPU 가속을 위한 CUDA 및 cuDNN 활용 [65], 실시간 비디오 분석 및 처리를 위한 DeepStream SDK 활용 [69] 등의 최적화 기법들을 고려해야 한다.

알고리즘 및 기술의 연산 요구 사항과 실제 하드웨어에서의 성능을 예측하기 위해서는 MLPerf 벤치마크 결과 [66]를 참고할 수 있다. MLPerf는 다양한 딥러닝 모델에 대한 Jetson AGX Orin의 성능 지표 (예: YOLOv8n FPS)를 제공하며, 이를 통해 모델 선택 및 성능 예측에 유용한 정보를 얻을 수 있다. 또한, 커뮤니티 및 개발자 포럼 정보 [71]는 실제 사용 사례 및 성능 경험을 공유하므로, 실제 시스템 구축 시 발생할 수 있는 문제점들을 파악하고 해결 방안을 모색하는 데 도움이 된다.

Jetson AGX Orin은 로봇 파지 작업에 필요한 고성능 컴퓨팅 능력을 제공하지만, 실시간 성능을 달성하기 위해서는 모델 최적화 및 하드웨어 가속 기술을 적극적으로 활용해야 한다. MLPerf 벤치마크 결과는 모델 선택 및 성능 예측에 유용한 정보를 제공하며, 개발자 포럼은 실제 사용 경험을 통해 얻은 valuable insights를 제공한다. 따라서 Jetson AGX Orin 환경에서 실시간 Unlabeled 객체 파지 시스템을 구축하기 위해서는 효율적인 알고리즘 선택과 함께 하드웨어 특성에 맞는 최적화 전략이 필수적이며, 모델 경량화, TensorRT 활용, GPU 가속 등을 통해 실시간 성능을 확보하고, 실제 하드웨어에서의 성능을 예측하기 위해 벤치마크 결과 및 커뮤니티 정보를 참고해야 한다.

3. 자연어 명령 기반 로봇 파지 연구 **(VLA)**

3.1 비전**-**언어**-**행동 **(VLA)** 모델의 최신 연구 동향

비전-언어-행동 (Vision-Language-Action, VLA) 모델은 시각적 입력과 자연어 명령을 이해하고 로봇 행동을 생성하는 foundation model로 [75], 로봇이 인간의 의도를 파악하고 다양한 작업을 수행할 수 있도록 하는 핵심 기술로 주목받고 있다. Google DeepMind에서 개발한 RT-2는 이러한 VLA 모델의 대표적인 예시이다.[75] VLA 모델은 대규모 인터넷 또는 시뮬레이션 데이터로 사전 학습하여 task-specific 로봇 데이터에 대한 의존성을 줄이고 [76], "빨간색 블록을 집어 들어"와 같은 고수준 명령을 해석하고 실행할 수 있으며 [76], 다양한 작업, 객체, 로봇 플랫폼에 대한 지식 이전을 가능하게 한다.[76] 또한, 인식, 작업 이해, 제어를 단일 모델에서 공동으로 처리하여 시스템 설계 단순화 및 robust성을 향상시킬 수 있다는 장점을 가진다.[76]

하지만 VLA 모델은 여전히 해결해야 할 과제와 한계점을 가지고 있다. 특정 환경 또는 작업에서 학습된 모델은 새로운 환경 또는 작업으로 일반화하는 데 어려움을 겪을 수 있으며 [77], 모델 아키텍처의 복잡성과 학습 및 추론의 확장성 또한 중요한 문제이다.[77] 다양하고 고품질의 로봇 데이터를 수집하는 것 역시 VLA 모델 연구의 주요 과제 중 하나이며 [77], 현재 VLA 모델은 주로 다음 단계 행동 예측에 집중하여 장기 계획이 필요한 작업에서는 어려움을 겪을 수 있다.[76]

3.2 다양한 **VLA** 모델의 작동 방식 및 특징 비교 분석

다양한 작동 방식과 특징을 가진 VLA 모델들이 활발하게 연구되고 있으며, 각 모델은 특정 강점과 응용 분야에 초점을 맞추고 있다.

DexGraspVLA는 사전 학습된 VLM을 고수준 작업 계획자로 활용하고, diffusion 기반 정책을 저수준 행동 제어기로 사용하여 일반적인 손재주 파지를 달성하는 계층적 프레임워크이다.[78] 이 모델은 다양한 시각적 및 언어적 입력을 도메인 불변 표현으로 변환하여 imitation learning의 효율성을 높이고, 수천 가지의 보지 못한 객체, 조명, 배경 조합에서 높은 손재주 파지 성공률을 보인다.

CoT-VLA (Visual Chain-of-Thought Reasoning for VLAs)는 미래 이미지 프레임을 시각적 목표로 예측하여 명시적인 시각적 사고 과정을 VLA 모델에 통합하고, 이를 통해 복잡한 조작 작업 성능을 향상시킨다.[80] 이 모델은 시각적 사고 과정을 통해 VLA 모델의 추론 능력 및 장기 계획 가능성을 향상시키는 새로운 접근 방식을 제시한다.

SafeVLA는 시뮬레이션 환경에서 대규모 제약 학습을 통해 안전 제약 조건을 명시적으로 통합하여 환경, 로봇 하드웨어, 인간을 보호하는 것을 목표로 한다.[81] 이 모델은 안전성과 작업 성능 간의 균형을 효과적으로 유지하며, 기존 state-of-the-art 방법보다 향상된 결과를 보인다.

EF-VLA (Early Fusion VLA)는 CLIP의 시각-언어 이해 능력을 활용하여 시각 및 언어 특징을 초기 단계에서 융합하고, 작업 지침과 관련된 세분화된 시각-언어 토큰을 추출하여 정책 네트워크에 전달한다.[82] 이 모델은 fine-tuning 없이도 새로운 작업에 대한 뛰어난 일반화 능력을 입증한다.

Pi0는 flow matching을 사용하여 부드럽고 실시간적인 행동 궤적을 생성하며, 다양한 환경 및 로봇 형태에 대한 강력한 일반화 능력을 보유한 VLA 모델이다.[83] 이 모델은 행동 표현 방식의 중요성을 강조하며, flow matching 기반의 새로운 VLA 모델을 제시한다.

표 **2:** 다양한 **VLA** 모델의 특징 비교

모델 이름	핵심 접근 방식	주요 강점	주요 제한 사항/과제
DexGraspVLA	계층적 구조, VLM 계획자, Diffusion 제어기	뛰어난 일반화 능력, 다양한 환경 및 객체에 대한 robust한 파지 성능	
CoT-VLA	시각적 사고 과정 통합	복잡한 조작 작업 성능 향상, 추론 능력 및	

		장기 계획 가능성 향상	
SafeVLA	안전 제약 조건 명시적 통합	안전성 및 작업 성능 간 균형 유지, 실제 로봇 환경에서의 안전한 배포 가능성 제시	
EF-VLA	초기 시각-언어 특징 융합	뛰어난 일반화 능력, fine-tuning 없이 새로운 작업 수행 가능성	CLIP에 특화됨
Pi0	Flow Matching 기반 행동 궤적 생성	부드럽고 실시간적인 행동 궤적 생성, 다양한 환경 및 로봇 형태에 대한 강력한 일반화 능력	

4. 두 가지 연구 주제의 비교 분석 및 연구 방향 제시

4.1 Unlabeled 객체 파지 **vs. VLA** 기반 파지**:** 핵심 차이점 및 유사점

Unlabeled 객체 파지 연구는 특정 객체에 대한 사전 정보 없이 자연어 명령 (주로 속성 기반) 또는 시각적 특징을 기반으로 객체를 인식하고 파지하는 데 초점을 맞춘다. 이 접근 방식은 CLIP과 같은 VLM을 활용하여 zero-shot 인식을 수행하고, RGBD 데이터를 이용하여 파지 지점을 결정하는 경향이 있다.[4] 반면, VLA 기반 파지 연구는 시각적 입력과 자연어 명령을 통합적으로 이해하여 로봇의 행동 (파지 포함)을 직접 생성하는 데 초점을 맞춘다. 이 접근 방식은 end-to-end 학습 방식을 통해 복잡한 작업 수행 가능성을 제시한다.[75]

두 연구 주제 모두 자연어 이해를 기반으로 로봇이 환경과 상호작용하는 것을 목표로 하며, 궁극적으로 사용자의 의도를 로봇의 행동으로 연결하는 것을 지향한다는 유사점을 가진다. 또한, RGBD 카메라를 주요 센서로 활용하고, 딥러닝 모델을 핵심 기술로 사용하는 공통점이 있다. 그러나 Unlabeled 객체 파지는 객체 인식 및 파지 지점 결정을 위한 모듈화된 접근 방식을 취하는 반면, VLA 기반 파지는 end-to-end 방식으로 전체 파지 과정을 모델링한다는 뚜렷한 차이점을 보인다. VLA는 더 복잡한 작업 및 상황 이해에 유리할 수 있지만, 학습 데이터 요구량이 더 많을 수 있다.

4.2 각 접근 방식의 장단점 분석

Unlabeled 객체 파지 접근 방식은 모듈화된 구조로 각 구성 요소 (인식, 파지 계획)의 개발 및 개선이 용이하며, CLIP과 같은 사전 학습된 모델을 활용하여 학습 데이터 요구량이

상대적으로 적고, 속성 기반 접근 방식을 통해 새로운 객체에 대한 일반화 능력이 우수하다는 장점을 가진다. 하지만 복잡한 작업 흐름 또는 다단계 추론이 필요한 경우 어려움이 발생할 수 있으며, 자연어 명령의 ambiguity 처리 및 상황 맥락 이해에 한계가 존재할 수 있다.

반면, VLA 기반 파지 접근 방식은 end-to-end 학습을 통해 복잡한 작업 및 상황에 대한 통합적인 이해가 가능하며, 자연어 명령과 시각적 정보를 직접적으로 연결하여 더 자연스러운 로봇 제어 가능성을 제시하고, 장기 계획 및 추론 능력 발전 가능성을 내포한다. 그러나 학습 데이터 요구량이 매우 많고, 모델 해석 및 디버깅이 어려우며, 새로운 환경 또는 작업에 대한 일반화 어려움이 발생할 수 있고, 실시간 성능 확보를 위한 모델 경량화 및 최적화가 필수적이라는 단점을 가진다.

4.3 사용자의 연구 목표에 따른 최적 연구 방향 제시

사용자의 연구 목표가 특정 속성을 가진 다양한 미지의 객체를 파지하는 데 초점을 맞춘다면, Unlabeled 객체 파지 연구가 더 적합할 수 있다. 특히 CLIP과 속성 기반 파지 방법을 결합하고, RGBD 데이터를 활용하여 정확한 파지 지점을 결정하는 연구 방향을 고려해 볼 수 있다. 사용자의 현재 지식 수준 (CLIP에 대한 이해)과 관심사를 고려할 때, 첫 번째 연구 주제를 시작점으로 삼아 VLA 기반 파지로 점진적으로 확장하는 것도 좋은 전략이 될 수 있다. Unlabeled 객체 파지 연구를 통해 객체 인식 및 기본적인 파지 기술을 먼저 확보하고, VLA 모델 연구를 통해 더 복잡한 자연어 명령 처리 및 작업 수행 능력을 개발하는 단계적인 접근 방식을 취할 수 있다.

5. Isaac Sim 환경에서의 강화 학습 기반 파지 학습

5.1 Isaac Sim을 활용한 로봇 팔 파지 학습 방법

Isaac Sim은 NVIDIA에서 제공하는 로봇 시뮬레이터로, 물리적으로 정확한 환경에서 합성 데이터를 생성할 수 있으며, GPU 병렬 연산을 통해 빠른 시뮬레이션 및 학습을 지원한다.[84] Isaac Sim을 활용한 로봇 팔 파지 학습은 로봇 팔, 객체, 작업 공간 모델링, 센서 데이터 (RGB, Depth) 시뮬레이션, 파지 성공/실패 조건 정의 등의 단계를 포함하는 강화 학습 환경 구축을 통해 이루어진다.[85] 또한, 파지 성공, 안정성, 효율성 등을 반영하는 보상 함수를 설계하고 [93], 인간의 파지 시연 데이터를 모방하여 초기 정책을 학습시키는 Imitation Learning [86]이나 다양한 객체 및 환경에 대한 일반화 능력을 향상시키는 Meta-Learning [87] 등의 방법들을 활용할 수 있다.

5.2 SAC 및 **PPO** 알고리즘을 이용한 **Continuous Action Space** 학습 전략

SAC (Soft Actor-Critic)는 Off-policy 알고리즘으로, exploration과 exploitation 간의 균형을 효과적으로 유지하며 continuous action space에서 좋은 성능을 보인다.[94] PPO (Proximal Policy Optimization)는 On-policy 알고리즘으로, policy 업데이트 시 안정성을 확보하여 학습 효율성을 높이며, continuous action space에도 적용 가능하다.[95]

Continuous action space는 로봇 팔의 joint 각도 또는 end-effector의 6DoF (위치 및 방향)로 표현될 수 있으며 [95], 학습 안정성 및 효율성 향상을 위해 action space 정규화 및 적절한 탐험 전략 (예: Gaussian noise 추가)이 필요하다.

5.3 강화 학습 적용 시 고려 사항 및 구체적인 사례

강화 학습을 실제 로봇 시스템에 적용하기 위해서는 시뮬레이션에서 학습된 정책을 실제 로봇에 적용할 때 발생하는 성능 저하 문제인 Sim-to-Real Transfer 문제를 고려해야 한다.[96] 이러한 문제를 해결하기 위해 시뮬레이션 환경의 다양한 파라미터 (조명, 텍스처, 객체 속성 등)를 랜덤하게 변화시켜 실제 환경에서의 일반화 능력을 향상시키는 Domain Randomization [96], 실제 로봇의 동적 모델을 파악하여 시뮬레이션 환경에 반영함으로써 sim-to-real gap을 감소시키는 System Identification [97], 학습 시에는 추가적인 정보 (예: 객체의 정확한 포즈)를 활용하고, 테스트 시에는 실제 센서 데이터만 사용하는 Privileged Learning [97] 등의 기법들을 활용할 수 있다.

6. 결론

본 보고서는 자연어 기반 로봇 파지 연구의 두 가지 주요 방향인 Unlabeled 객체 파지와 VLA 기반 파지에 대한 최신 연구 동향을 분석하고, 각 접근 방식의 특징과 장단점을 비교 분석하였다. Unlabeled 객체 파지는 CLIP과 같은 VLM의 zero-shot 능력과 속성 기반 파지 방법을 활용하여 새로운 객체를 인식하고 파지하는 데 초점을 맞추고 있으며, VLA 기반 파지는 시각적 입력과 자연어 명령을 통합적으로 이해하여 로봇의 파지 행동을 직접 생성하는 end-to-end 학습 방식을 지향한다. 두 연구 방향 모두 자연어 이해를 기반으로 로봇이 환경과 상호작용하는 것을 목표로 하지만, 접근 방식, 장단점, 그리고 필요한 기술적 고려 사항에서 차이를 보인다. 또한, Isaac Sim 환경에서의 강화 학습을 통한 로봇 팔 파지 학습 방법과 SAC 및 PPO 알고리즘의 적용 전략, 그리고 Sim-to-Real Transfer 문제 해결을 위한 다양한 기법들을 살펴보았다. 사용자의 연구 목표와 현재 지식 수준을 고려할 때, Unlabeled 객체 파지 연구를 시작으로 점진적으로 VLA 기반 파지로 확장하는 연구 방향을 제안한다. 향후 연구에서는 Unlabeled 객체 파지의 viewpoint 및 occlusion robust성 향상, 복잡한 자연어 명령 처리 능력 향상, 그리고 VLA 기반 파지의 일반화 능력 향상, 안전성 확보, 장기 계획 능력 개발 등에 대한 심층적인 연구가 필요할 것으로 예상된다. 또한, Isaac Sim을 활용한 강화 학습 기반 파지 연구에서는 Sim-to-Real Transfer 성능 향상 및 새로운 강화 학습 알고리즘 탐색 등이 중요한 연구 주제가 될 수 있을 것이다.

참고 자료

1. Zero-shot object detection - Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/docs/transformers/tasks/zero_shot_object_detection
2. Simple but Effective: CLIP Embeddings for Embodied AI - CVF Open Access, 4월 25, 2025에 액세스, https://openaccess.thecvf.com/content/CVPR2022/papers/Khandelwal_Simple_bu

t_Effective_CLIP_Embeddings_for_Embodied_AI_CVPR_2022_paper.pdf
3. Active Open-Vocabulary Recognition: Let Intelligent Moving Mitigate CLIP Limitations, 4월 25, 2025에 액세스, https://openaccess.thecvf.com/content/CVPR2024/papers/Fan_Active_Open-Vocabulary_Recognition_Let_Intelligent_Moving_Mitigate_CLIP_Limitations_CVPR_2024_paper.pdf
4. Attribute-Based Robotic Grasping with Data-Efficient Adaptation - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2501.02149
5. Attribute-Based Robotic Grasping with Data-Efficient Adaptation - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2501.02149v1
6. Reliable pretrained object recognition models : r/computervision - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/computervision/comments/1aunyj5/reliable_pretrained_object_recognition_models/
7. RGB-D Object Recognition for Deep Robotic Learning - AMS Tesi di Laurea, 4월 25, 2025에 액세스, https://amslaurea.unibo.it/id/eprint/14537/1/tesi_cimmino_martin.pdf
8. CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection | OpenReview, 4월 25, 2025에 액세스, https://openreview.net/forum?id=QW5ouyylgG¬eId=PPzQ7xoXoO
9. CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection - NeurIPS, 4월 25, 2025에 액세스, https://proceedings.neurips.cc/paper_files/paper/2023/file/e352b765e625934ce86919995e2371aa-Paper-Conference.pdf
10. CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection NeurIPS 2023 - Yang Cao, 4월 25, 2025에 액세스, https://yangcaoai.github.io/publications/CoDA.html
11. [2310.02960] CoDA: Collaborative Novel Box Discovery and Cross-modal Alignment for Open-vocabulary 3D Object Detection - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/abs/2310.02960
12. [2408.13936] OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/abs/2408.13936
13. OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation | AI Research Paper Details - AIModels.fyi, 4월 25, 2025에 액세스, https://www.aimodels.fyi/papers/arxiv/opennav-efficient-open-vocabulary-3d-object-detection
14. OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation, 4월 25, 2025에 액세스, https://arxiv.org/html/2408.13936v1
15. OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/383428546_OpenNav_Efficient_Open_Vocabulary_3D_Object_Detection_for_Smart_Wheelchair_Navigation
16. Official code for the OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation - ACVR Workshop at ECCV'24 - GitHub, 4월 25,

2025에 액세스, https://github.com/EasyWalk-PRIN/OpenNav
17. Luca Tonin | Papers With Code, 4월 25, 2025에 액세스, https://paperswithcode.com/author/luca-tonin
18. Boosting grape bunch detection in RGB-D images using zero-shot annotation with Segment Anything and GroundingDINO, 4월 25, 2025에 액세스, https://hau.repository.guildhe.ac.uk/id/eprint/18166/1/Fernando%20Auat%20Cheein%20Boosting%20grape%20bunch%20detection%20in%20RGB-D%20images%20VoR%20OCR%20UPLOAD.pdf
19. Attribute-Based Robotic Grasping With Data-Efficient Adaptation | Request PDF, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/377372239_Attribute-Based_Robotic_Grasping_with_Data-Efficient_Adaptation
20. Yang Yang, 4월 25, 2025에 액세스, https://st2yang.github.io/
21. Publications | Choice Robotics Lab - University of Minnesota, 4월 25, 2025에 액세스, https://choice.umn.edu/publications
22. Attribute-Based Robotic Grasping with Data-Efficient Adaptation, 4월 25, 2025에 액세스, https://colab.ws/articles/10.1109%2Ftro.2024.3353484
23. Attribute-Based Robotic Grasping with Data-Efficient Adaptation. - dblp, 4월 25, 2025에 액세스, https://dblp.org/rec/journals/corr/abs-2501-02149.html
24. CLIP Alternatives: A Guide - Roboflow, 4월 25, 2025에 액세스, https://roboflow.com/model-alternatives/clip
25. Top 8 Alternatives to the Open AI CLIP Model - Encord, 4월 25, 2025에 액세스, https://encord.com/blog/open-ai-clip-alternatives/
26. CLIPort: What and Where Pathways for Robotic Manipulation - OpenReview, 4월 25, 2025에 액세스, https://openreview.net/forum?id=9uFiX_HRsIL
27. Robotic-CLIP: Fine-tuning CLIP on Action Data for Robotic Applications - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.17727v1
28. openaccess.thecvf.com, 4월 25, 2025에 액세스, https://openaccess.thecvf.com/content/CVPR2024/papers/Vuong_Language-driven_Grasp_Detection_CVPR_2024_paper.pdf
29. Reasoning Grasping via Multimodal Large Language Model - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2402.06798v1
30. Reasoning Grasping via Multimodal Large Language Model - GitHub, 4월 25, 2025에 액세스, https://raw.githubusercontent.com/mlresearch/v270/main/assets/jin25a/jin25a.pdf
31. Reasoning Grasping via Multimodal Large Language Model - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2402.06798v2
32. [2402.06798] Reasoning Grasping via Multimodal Large Language Model - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/abs/2402.06798
33. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach - Robotics, 4월 25, 2025에 액세스, https://www.roboticsproceedings.org/rss14/p21.pdf
34. GR-ConvNet v2: A Real-Time Multi-Grasp Detection Network for Robotic Grasping - PMC, 4월 25, 2025에 액세스, https://pmc.ncbi.nlm.nih.gov/articles/PMC9415764/

35. citeseerx.ist.psu.edu, 4월 25, 2025에 액세스, https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=3c104b0e182a5f514d3aebecc93629bbcf1434ac
36. Efficient grasping from RGBD images: Learning using a new rectangle representation, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/221072021_Efficient_grasping_from_RGBD_images_Learning_using_a_new_rectangle_representation
37. www.roboticsproceedings.org, 4월 25, 2025에 액세스, https://www.roboticsproceedings.org/rss09/p12.pdf
38. A New Robotic Grasp Detection Method based on RGB-D Deep Fusion* - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/363312173_A_New_Robotic_Grasp_Detection_Method_based_on_RGB-D_Deep_Fusion
39. Combining RGB and Points to Predict Grasping Region for Robotic Bin-Picking - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/pdf/1904.07394
40. dkanou.github.io, 4월 25, 2025에 액세스, https://dkanou.github.io/publ/C46.pdf
41. Optimized Grasp Pose Detection Using RGB Images for Warehouse Picking Robots - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.19494
42. Lan-grasp: Using Large Language Models for Semantic Object Grasping - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2310.05239v2
43. AffordDexGrasp: Open-set Language-guided Dexterous Grasp with Generalizable-Instructive Affordance - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2503.07360v1
44. Affordance Prediction Analysis - Dexterous Functional Grasping, 4월 25, 2025에 액세스, https://dexfunc.github.io/rebuttal/affordance.html
45. Lightweight Neural Networks for Affordance Segmentation: Enhancement of the Decoder Module | Request PDF - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/377368460_Lightweight_Neural_Networks_for_Affordance_Segmentation_Enhancement_of_the_Decoder_Module
46. Lightweight Language-driven Grasp Detection using Conditional Consistency Model - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2407.17967v1
47. Pick and Place — moveit_tutorials Noetic documentation - GitHub Pages, 4월 25, 2025에 액세스, https://moveit.github.io/moveit_tutorials/doc/pick_place/pick_place_tutorial.html
48. moveit_msgs/Grasp Message, 4월 25, 2025에 액세스, http://docs.ros.org/en/noetic/api/moveit_msgs/html/msg/Grasp.html
49. Pick and Place Tutorial — moveit_tutorials Kinetic documentation - ROS, 4월 25, 2025에 액세스, http://docs.ros.org/en/kinetic/api/moveit_tutorials/html/doc/pick_place/pick_place_tutorial.html
50. Obstacles Management in moveit - Google Groups, 4월 25, 2025에 액세스, https://groups.google.com/g/moveit-users/c/EI73skgnGVk
51. Perception Pipeline Tutorial — MoveIt Documentation: Humble documentation, 4월 25, 2025에 액세스, https://moveit.picknik.ai/humble/doc/examples/perception_pipeline/perception_pi

peline_tutorial.html
52. MoveIt Grasps, 4월 25, 2025에 액세스, https://moveit.picknik.ai/main/doc/examples/moveit_grasps/moveit_grasps_tutorial.html
53. GPD (Grasp Pose Detection) Library Integration · Issue #566 - GitHub, 4월 25, 2025에 액세스, https://github.com/moveit/moveit/issues/566
54. MoveIt Deep Grasps — MoveIt Documentation: Humble documentation, 4월 25, 2025에 액세스, https://moveit.picknik.ai/humble/doc/examples/moveit_deep_grasps/moveit_deep_grasps_tutorial.html
55. Grasp pose detection (GPD) and Dex-Net support · Issue #2188 - GitHub, 4월 25, 2025에 액세스, https://github.com/ros-planning/moveit/issues/2188
56. GSoC: Creating a default grasping library - MoveIt - ROS Discourse, 4월 25, 2025에 액세스, https://discourse.ros.org/t/gsoc-creating-a-default-grasping-library/8521
57. Getting Started with the MoveIt 2 Task Constructor - Automatic Addison, 4월 25, 2025에 액세스, https://automaticaddison.com/getting-started-with-the-moveit-2-task-constructor/
58. Pick and Place — MoveIt Documentation: Humble documentation, 4월 25, 2025에 액세스, https://moveit.picknik.ai/humble/doc/examples/pick_place/pick_place_tutorial.html
59. Pick and Place - MoveIt 2 Documentation - PickNik Robotics, 4월 25, 2025에 액세스, https://moveit.picknik.ai/main/doc/examples/pick_place/pick_place_tutorial.html
60. Concepts | MoveIt, 4월 25, 2025에 액세스, https://moveit.ai/documentation/concepts/
61. Perception Pipeline Tutorial — MoveIt Documentation - PickNik Robotics, 4월 25, 2025에 액세스, https://moveit.picknik.ai/main/doc/examples/perception_pipeline/perception_pipeline_tutorial.html
62. Create a Pick and Place Task Using MoveIt 2 and Perception - Automatic Addison, 4월 25, 2025에 액세스, https://automaticaddison.com/create-a-pick-and-place-task-using-moveit-2-and-perception/
63. Perception Pipeline Tutorial — moveit_tutorials Kinetic documentation - ROS 2, 4월 25, 2025에 액세스, http://docs.ros.org/en/kinetic/api/moveit_tutorials/html/doc/perception_pipeline/perception_pipeline_tutorial.html
64. MoveIt Tutorials — moveit_tutorials Noetic documentation - GitHub Pages, 4월 25, 2025에 액세스, https://moveit.github.io/moveit_tutorials/
65. What is Jetson Agx Orin? - sinsmart industrial pc computer, 4월 25, 2025에 액세스, https://www.sinsmarts.com/blog/what-is-jetson-agx-orin/
66. NVIDIA Jetson AGX Orin and RTX 4090 Comparison - Lowtouch.Ai, 4월 25,

2025에 액세스, https://www.lowtouch.ai/nvidia-jetson-agx-orin-and-rtx-4090-in-ai-applications/
67. Deploying DeepSeek AI on NVIDIA Jetson AGX Orin: A Free, Open-Source MIT-Licensed Solution for High-Performance Edge AI in Natural Language Processing and Computer Vision - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/388401833_Deploying_DeepSeek_AI_on_NVIDIA_Jetson_AGX_Orin_A_Free_Open-Source_MIT-_Licensed_Solution_for_High-Performance_Edge_AI_in_Natural_Language_Processing_and_Computer_Vision
68. Build an AI-driven object detection algorithm with balenaOS & alwaysAI - balena Blog, 4월 25, 2025에 액세스, https://blog.balena.io/build-an-ai-driven-object-detection-algorithm-with-balenaos-and-alwaysai/
69. Tutorial: Real-Time Object Detection with DeepStream on Nvidia Jetson AGX Orin, 4월 25, 2025에 액세스, https://thenewstack.io/tutorial-real-time-object-detection-with-deepstream-on-nvidia-jetson-agx-orin/
70. Jetson Benchmarks - NVIDIA Developer, 4월 25, 2025에 액세스, https://developer.nvidia.com/embedded/jetson-benchmarks
71. Real-Time Tool Detection and Tracking with Jetson : r/computervision - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/computervision/comments/1h9dpfe/realtime_tool_detection_and_tracking_with_jetson/
72. Training model, object detetion - Jetson Orin NX - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/training-model-object-detetion/265711
73. Image recognition and object grabbing with robotic arms - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/image-recognition-and-object-grabbing-with-robotic-arms/212781
74. Sdk 4.1 on jetson agx orin jp 6.0 - Stereolabs Forums, 4월 25, 2025에 액세스, https://community.stereolabs.com/t/sdk-4-1-on-jetson-agx-orin-jp-6-0/6433
75. Vision-language-action model - Wikipedia, 4월 25, 2025에 액세스, https://en.wikipedia.org/wiki/Vision-language-action_model
76. Vision-Language-Action (VLA) Models: LLMs for robots, 4월 25, 2025에 액세스, https://www.blackcoffeerobotics.com/blog/vision-language-action-vla-models-llms-for-robots
77. Survey on Vision-Language-Action Models - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2502.06851v1
78. DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping, 4월 25, 2025에 액세스, https://dexgraspvla.github.io/assets/paper/DexGraspVLA.pdf
79. DexGraspVLA: A Vision-Language-Action Framework Towards General Dexterous Grasping, 4월 25, 2025에 액세스, https://dexgraspvla.github.io/
80. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action

Models - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2503.22020v1
81. SafeVLA: Towards Safety Alignment of Vision-Language-Action Model via Safe Reinforcement Learning - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2503.03480v1
82. Early Fusion Helps Vision Language Action Models Generalize Better - OpenReview, 4월 25, 2025에 액세스, https://openreview.net/forum?id=KBSHR4h8XV
83. π0 and π0-FAST: Vision-Language-Action Models for General Robot Control, 4월 25, 2025에 액세스, https://huggingface.co/blog/pi0
84. An Easy to Use Deep Reinforcement Learning Library for AI Mobile Robots in Isaac Sim, 4월 25, 2025에 액세스, https://www.mdpi.com/2076-3417/12/17/8429
85. Grasp Editor - Isaac Sim Documentation, 4월 25, 2025에 액세스, https://docs.isaacsim.omniverse.nvidia.com/latest/robot_setup/grasp_editor.html
86. Teleoperation and Imitation Learning — Isaac Lab Documentation, 4월 25, 2025에 액세스, https://isaac-sim.github.io/IsaacLab/main/source/overview/teleop_imitation.html
87. YitianShi/MetaIsaacGrasp: Isaaclab-based grasp learning test bench - GitHub, 4월 25, 2025에 액세스, https://github.com/YitianShi/MetaIsaacGrasp
88. Grasp Editor — Isaac Sim 4.2.0 (OLD) - NVIDIA Omniverse, 4월 25, 2025에 액세스, https://docs.omniverse.nvidia.com/isaacsim/latest/advanced_tutorials/tutorial_grasp_editor.html
89. [Isaac Sim Tutorial - Core API] Introduction to Lecture Code Setup - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=hP01sOYtXKM
90. Manipulation — isaac_ros_docs documentation - NVIDIA Isaac ROS, 4월 25, 2025에 액세스, https://nvidia-isaac-ros.github.io/concepts/manipulation/index.html
91. Reference Architecture — Isaac Lab Documentation - GitHub Pages, 4월 25, 2025에 액세스, https://isaac-sim.github.io/IsaacLab/main/source/refs/reference_architecture/index.html
92. Arm Manipulation in Isaac Sim Using MoveIt Servo - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=-jqepxgzs_I
93. CLIP-Motion: Learning Reward Functions for Robotic Actions Using Consecutive Observations - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2311.03485v2
94. Wheeled Robot Setup and Reinforcement Learning Using Isaac Sim - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=rVCHgXI9E6o
95. IsaacGymEnvs/docs/rl_examples.md at main - GitHub, 4월 25, 2025에 액세스, https://github.com/isaac-sim/IsaacGymEnvs/blob/main/docs/rl_examples.md
96. Reinforcement Learning for Humanoid Robot with Zero-Shot Sim2Real Transfer - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2404.05695v2
97. Sim-to-Real Gap in RL: Use Case with TIAGo and Isaac Sim/Gym - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/387578217_Sim-to-Real_Gap_in_RL_Use_Case_with_TIAGo_and_Isaac_SimGym