Jetson AGX Orin 기반 nanoOWL과 nanoSAM 결합 로봇 그래스핑 최신 연구 동향 (2023 이후)

1. Jetson AGX Orin에서 nanoOWL + nanoSAM 결합 실시간 그래스핑 시스템

Jetson AGX Orin과 같은 엣지 AI 플랫폼에서 **nanoOWL**과 **nanoSAM**을 결합해 자연어로 지목한 물체를 세그멘테이션하고 파지(把持)하는 시스템이 등장했습니다. **nanoOWL**은 OWL-ViT(Open-Vocabulary Object Detection 모델)을 TensorRT로 최적화해 Jetson Orin에서 실시간으로 동작하게 한 것이고, **nanoSAM**은 Segment Anything Model(SAM)을 경량화(distillation)하여 Orin에서 실시간 인퍼런스를 가능하게 한 변형 모델입니다 (GitHub - NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT.) (NVIDIA-AI-IOT/nanosam: A distilled Segment Anything (SAM) model ...). 두 모델을 조합하면 사용자가 입력한 **자연어 객체 명칭**에 따라 해당 물체를 검출하고 세그멘테이션 마스크를 생성하여, 로봇이 원하는 물체를 파지할 수 있는 **제로샷 개방형 인스턴스 세그멘테이션** 파이프라인을 구현할 수 있습니다 (GitHub - NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT.). 이러한 개념은 **개방형 세계(Open-World) 그래스핑** 연구와 연결되어, 한정된 카테고리 외의 임의 사물에 대한 파지를 목표로 합니다.

- **개방형 세계 그래스핑(OWG) 파이프라인**: Tziafas 등(CoRL 2024)은 대규모 비전-언어 모델(VLM)과 세그멘테이션 및 파지 합성모델을 결합한 **OWG(Open-World Grasping)** 파이프라인을 제안했습니다 (Towards Open-World Grasping with Large Vision-Language Models). 이 시스템은 **자연어 지시**(예: *"아이들이 갖고 놀 만한 것을 집어줘"*)를 받아, 세그멘테이션 모델(SAM 기반)로 장면의 모든 객체 마스크를 생성하고 각 객체에 ID를 부여한 이미지를 VLM에 입력합니다 (Academic Project Page). VLM은 시각-언어 추론을 통해 지시된 목표 물체를 식별하고, 주변 장애물은 치워야 하는지 판단한 뒤, 파지 합성 모델을 호출해 후보 파지 자세들을 생성합니다. 마지막으로 물체의 형상과 주변 접촉 정보를 고려한 **그립 평가(ranking)**를 통해 최적 파지 자세를 선택합니다 (Academic Project Page) (Academic Project Page). 이 방식은 **클러터(clutter)** 속 개방형 지시어에 대한 견고한 시각적 그라운딩과 접촉-기반 판단을 가능하게 하여, 이전의 LLM 기반 로봇계획보다 **높은 성공률**을 보였습니다 (Academic Project Page). 해당 연구의 소스코드가 공개되어 있습니다.

- **비전-언어-액션 공동 모델링**: Xu 등(ICRA 2023)은 **언어 지시 기반 목표-지향 그래스핑**을 위해 **비전-언어-행동의 공동 학습 모델**을 제안하였습니다 (GitHub - xukechun/Vision-Language-Grasping: [ICRA 2023] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter). 기존 방법들은 **시각적 그라운딩** (예: 물체 인식)과 **파지**를 별도 단계로 수행하면서, 사전 정의된 객체 레이블이나 속성에 의존해야 했습니다. 반면 이 연구는 **객체 중심 표현(object-centric representation)**으로 언어, 시각, 행동을 통합하여, 별도의 규칙 설계 없이 자유로운 자연어 명령을 처리합니다 (GitHub - xukechun/Vision-Language-Grasping: [ICRA 2023] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter) (GitHub - xukechun/Vision-Language-Grasping: [ICRA 2023] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter). 미리 학습된 멀티모달 모델(예: CLIP)과 파지 모델의 강력한 프라이어를 활용하여 샘플 효율을 높이고 시뮬레이션-실세계 간 갭을 줄였으며, 시뮬레이션 및 실제 로봇 실험에서 **적은 동작 횟수로 더 높은 작업 성공률**을 달성했습니다 (GitHub - xukechun/Vision-Language-Grasping: [ICRA 2023] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter). 또한 보이지 않은 새로운 객체나 새로운 언어 명령에도 범용적으로 잘 일반화되는 것을 보여주었습니다. 해당 연구의 공식 구현 코드도 GitHub에 공개되어 있습니다 ([2302.12610] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter).

- **자연어+세그멘테이션 기반 grasp 검출**: Vo 등(IROS 2024 예정)은 **마스크-guided 어텐션** 방식을 통해 **자연어 지시 기반 그립 검출** 성능을 향상시켰습니다 (Language-driven Grasp Detection with Mask-guided Attention). 세그멘테이션 마스크로부터 추출한 특성을 비전 트랜스포머의 어텐션에 통합하여, 시각정보와 **텍스트 명령**을 함께 처리함으로써 복잡한 환경의 가려짐(occlusion) 상황에서도 정확도를 높였습니다 (Language-driven Grasp Detection with Mask-guided Attention). 제안된 방법은 **기존 방법 대비 명확히 향상된 grasp 성공 점수**를 달성했고, 실제 로봇 실험을 통해 유효성을 검증했습니다 (Language-driven Grasp Detection with Mask-guided Attention). 이처럼 최첨단 연구들은 Jetson Orin 상의 nanoOWL+nanoSAM 조합과 유사하게 **개방형 어휘**의 객체인식과 세그멘테이션을 결합하여, 사람의 언어 명령으로 특정 물체를 인식하고 잡을 수 있는 시스템 구현 가능성을 열어가고 있습니다.

2. 세그멘테이션 마스크 + Depth + 강화학습을 활용한 그래스핑

세그멘테이션 마스크와 깊이 정보를 결합하여 강화학습(RL)으로 로봇 파지 동작을 학습시키는 접근은, 특히 **SAC(Soft Actor-Critic)**와 같은 Off-policy RL 알고리즘의 활용으로 주목받고 있습니다. 이러한 방법들은 **이미지 기반 고차원 상태 공간**에서도 효율적인 학습을 위해 **시각 피처**를 제공하는 세그멘테이션과, **물체의 3D 형상**을 제공하는 깊이 데이터를 활용합니다. 2023년 이후 여러 연구들이 **세분화된 시각 정보**를 정책학습에 통합하여 파지 성공률을 높이는 사례를 보고하고 있습니다.

- **Grasp-Anything (TAPG)**: Mosbach 등(2023)은 **Teacher-Augmented Policy Gradient** 기법으로 단계적으로 RL 정책을 학습하여, 최종적으로 **SAM 세그멘테이션 모델**을 활용한 임의 물체 파지를 시도하는 **Grasp Anything** 시스템을 발표했습니다 (Grasp Anything: Combining Teacher-Augmented Policy Gradient Learning with Instance Segmentation to Grasp Arbitrary Objects). 1단계에서는 객체 **포즈** 정보만으로 교사 정책을 학습하고, 2단계에서는 이를 **학생 비전 정책**으로 지도로(distillation) 삼아 세그멘테이션 기반 센서모달 정책을 학습합니다 (Grasp Anything: Combining Teacher-Augmented Policy Gradient Learning with Instance Segmentation to Grasp Arbitrary Objects). 이렇게 함으로써 고차원 영상 입력의 학습 난이도를 극복하면서도, **SAM으로부터 얻은 분할 마스크**를 활용해 **특정 객체**를 집어내는 정책을 습득합니다. 이 방법은 시뮬레이션에서 학습한 후 추가 튜닝 없이 **SAM의 프롬프트 세그멘테이션**을 이용해 실제 로봇에 **제로샷 이식(zero-shot transfer)**되었으며, 인간이 이해하기 쉬운 프롬프트로 다양한 객체를 클러터 속에서 집어내는 데 성공했습니다 (Grasp Anything: Combining Teacher-Augmented Policy Gradient Learning with Instance Segmentation to Grasp Arbitrary Objects). 실제 새로운 객체들에 대해서도 강인한 일반화 성능을 보여주고 있습니다. (실험 영상과 코드도 공개됨)

- **세그멘테이션+SAC 기반 Push-Grasp 협조**: Gao 등(2024)은 강화학습을 이용해 **복잡한 환경에서 목표 물체 파지 전 불필요한 물체를 치우는(push) 동작과 파지 동작을 협업**시키는 프레임워크를 제안하였습니다. 특히 **개선된 SAC 기반** 심층 강화학습 구조를 통해 **목표 객체 분할(segmentation)**과 연계된 정책을 학습함으로써, **단일 관측뷰의 Depth 정보**로부터 **충돌 없는 6-자유도 파지 자세**를 생성합니다. 이 연구는 수중 환경의 복잡한 장면을 배경으로, 로봇이 카메라 한 장면에서 분할된 목표를 인지하고 **푸시-그립 연속 제어**를 배우도록 했으며, 그 결과 기존 방식 대비 높은 성공률을 보였습니다. (IEEE TIM 2024, 코드 비공개지만 주요 알고리즘 제시)

- **CLIP 기반 시각-언어 강화학습**: Yang 등(2024, Ground4Act) 연구에서는 클러터된 장면에서 **특정 물체를 집기 위한 행동**을 **두 단계**로 학습시켰습니다 (Ground4Act: Leveraging visual-language model for collaborative pushing and grasping in clutter | OpenReview). 먼저 **시각-언어 모델**(예: CLIP)의 임베딩을 활용해 자연어로 지정된 **타겟 객체의 시각적 위치**를 인식하고, **DQN 기반의 정책**으로 **비타겟 물체를 치워내는** 행동을 학습시킵니다 (Ground4Act: Leveraging visual-language model for collaborative pushing and grasping in clutter | OpenReview). 이후 목표물이 잘 드러나면 로봇이 파지하도록 하는 것으로, Push와 Grasp를 통합한 **협동 정책**을 구축했습니다. 이 강화학습 프레임워크는 시뮬레이션과 실제에서 동일한 형식으로 동작하며, 사람의 언어 명령으로 **클러터 속 원하는 물체를 성공적으로 파지**하는 능력을 보

여주었습니다 (Ground4Act: Leveraging visual-language model for collaborative ...). (코드는 공개 (GitHub - HDU-VRLab/Ground4Act: This repository contains the implementation of Ground4Act, a two-stage approach for collaborative pushing and grasping in clutter using a visual-language model.))

이처럼 세그멘테이션을 통해 **상태공간을 구조화**하고 Depth로 물체의 3차원 정보를 보강함으로써, 강화학습 에이전트가 **효과적인 파지 정책**을 학습하도록 돕는 연구들이 활발합니다. 특히 **SAC**과 같이 연속적 제어에 강점이 있는 알고리즘과 결합하여, **안전하고 최적화된 파지 동작**(예: 충돌 회피) 학습에 적용하는 사례가 증가하고 있습니다. 이러한 접근은 시뮬레이션-현실 간 격차를 줄여 실환경 적용 가능성을 높이고 있습니다.

3. 객체 형상 특성을 고려한 최적 파지 포인트 및 접근 방향 결정 알고리즘

물체의 **기하학적 형상 정보**를 적극 활용하여 **이상적인 파지 지점(grasp point)**과 **접근 경로(approach direction)**를 결정하는 알고리즘도 진화하고 있습니다. 2023년 이후의 연구들은 6-DoF 파지 탐색에서 **물체의 표면 형태, 모서리, 부품 구조** 등을 고려해 파지 안정성을 높이거나, 특정 기능적 파지를 가능케 하는 방향을 찾는 데 주력합니다.

- **SE(3) 등변(equivariant) 그립 학습**: Hu 등(CoRL 2024)의 **OrbitGrasp**는 포인트클라우드로부터 각 후보 접촉 지점마다 **연속적인 접근 방향에 대한 그립 품질 함수를 학습**하는 프레임워크입니다 (OrbitGrasp: SE(3)-Equivariant Grasp Learning). 구체적으로, **구면상의 모든 방향(S^2)**에서 그립 성공도를 예측하도록 모델을 구성하여, 기존에 제한된 샘플링에 의존하던 방법보다 **더 높은 정확도와 효율**을 달성했습니다 (OrbitGrasp: SE(3)-Equivariant Grasp Learning). 이 모델은 $\mathrm{SE}(3)$-등변 성질을 갖도록 설계되어, 물체의 자세 변화에 강인하며 포인트클라우드 입력으로 학습합니다. UNet 스타일의 인코더-디코더를 통해 많은 포인트도 처리 가능하게 하여, 시뮬레이션 및 실제 로봇 실험에서 **기존 대비 뛰어난 파지 성공률**을 보였습니다 (OrbitGrasp: SE(3)-Equivariant Grasp Learning). (공식 구현 코드 공개 ([CoRL 2024] OrbitGrasp: SE(3)-Equivariant Grasp Learning - GitHub))

- **기능적 파지를 위한 접근 히트맵**: Aburub 등(2024)은 로봇 손가락으로 도구의 버튼이나 방아쇠 같은 **기능부**를 조작하기 위해, 물체 표면에 **접근 히트맵(approach heatmap)**을 생성하는 **Functional Eigen-Grasping** 기법을 발표했습니다 (Functional Eigen-Grasping Using Approach Heatmaps). 다지 손(multi-fingered hand)의 특정 "기능 손가락"을 지정하면, 해당 손가락이 도구의 기능부를 정확히 눌러줄 수 있는 **최적의 손바닥 배치 위치**들을 물체 표면 위 히트맵으로 표시합니다 (Functional Eigen-Grasping Using Approach Heatmaps). 이때 로봇의 **방향별 조작 용이도 지표(방향 매니퓰러빌리티)**를 활용하여 각 지점의 점수를 산정하고, 가장 높은 점수 지점으로 손바닥을 가져가도록 합니다 (Functional Eigen-Grasping Using Approach Heatmaps) (Functional Eigen-Grasping Using Approach Heatmaps). 그렇게 찾은 자세에서 **eigen-grasp**(주성분 그립 형태)를 사용하여 나머지 손가락들로 물체를 안정적으로 잡도록 합니다 (Functional Eigen-Grasping Using Approach Heatmaps). 이 방법은 **인간 시演이나 사전 데이터 없이** 다양한 크기와 디자인의 도구에 적용 가능하며, Shadow Hand(인간형 손)뿐 아니라 Barrett Hand(비인간형 손)에도 확장될 수 있음을 실험으로 보였습니다 (Functional Eigen-Grasping Using Approach Heatmaps). 즉, 물체의 **기능적 형상**(버튼 위치 등)을 고려한 최적 접근 방향을 자동으로 찾아주는 일반적인 해결책을 제시한 것입니다.

- **형상 인지 기반 6-자유도 파지**: 고전적으로는 물체의 **곡률이 낮은 평면 부분에 수직으로 접근**하거나, **모서리(edge)**를 물도록 그리퍼를 배치하는 전략이 널리 쓰였습니다. 최신 딥러닝 기반 6-DoF 파지 모델들도 이러한 **형상 특징**을 학습적으로 반영합니다. 예를 들어, Wang 등(2023)은 6-DoF 파지검출 네트워크에 **표면 법선 방향** 정보를 부여하여, 그리퍼가 물체 표면에 최대한 수직에 가깝게 접근하도록 유도하였습니다. 또한 물체의 **대칭성과 연장선**을 고려해 양쪽 집게가 안정적으로 걸칠 수 있는 지점을 찾는 알고리즘도 제안됩니다. 이런 기법들은 물체 형상이 파지 성공률에 미치는 영향을 정량화하여, **형상적 최적 조건**(예: 최대 접촉면 확보, 마찰각 최적화)을 만족하는 grasp pose를 선정합니다. 최근 리뷰 논문들

에 따르면, **포인트클라우드 기반** 알고리즘들이 물체 완전한 3D 형상을 활용하여 이러한 접근 방향 최적화를 수행하며, 기존 2D 기반 방법보다 복잡한 형상의 물체에도 우수한 성능을 보이고 있습니다 (OrbitGrasp: SE(3)-Equivariant Grasp Learning) (OrbitGrasp: SE(3)-Equivariant Grasp Learning).

요약하면, 최신 연구에서는 단순히 데이터에 의존한 그립 점 예측을 넘어, **물체의 형태적 이해**를 바탕으로 **파지 자세 검색 공간을 효과적으로 좁히고 최적화**하고 있습니다. 이는 특히 **복잡한 형상이나 도구적 기능을 가진 객체**의 조작에 필수적이며, 이러한 알고리즘으로 로봇은 보다 인간 수준의 섬세한 파지 동작을 수행하게 됩니다.

4. 복잡한 환경에서 그래스핑 성공률 향상 방법론

로봇이 **어수선한 환경**에서 물체를 집어낼 때 직면하는 도전으로는, **물체들 간의 밀집과 상호 간섭**, **부분 가려짐 (occlusion)**, 잘못된 인식 등이 있습니다. 2023년 이후 연구들은 이러한 복잡한 환경에서도 파지 성공률을 높이기 위한 다양한 전략을 모색하고 있습니다. 주된 방향으로는 **전략적 주변 물체 정리**, **강인한 인스턴스 분할 및 추적**, **다중 단계 의사결정** 등이 있습니다.

- **푸쉬-그립 결합**: 복잡한 환경에서 **목표 물체가 다른 객체들에 가려지거나 잡기 어려운 위치**에 있는 경우, 로봇이 단순 파지 시도를 하기보다는 먼저 환경을 정리하는 행동이 필요합니다. Ground4Act 등 앞서 소개된 연구에서는 **강화학습**을 통해 **비목표물 밀쳐내기(push)**와 **목표물 파지(grasp)**를 통합했습니다 (Ground4Act: Leveraging visual-language model for collaborative pushing and grasping in clutter | OpenReview). 이처럼 **밀치기-집기**의 **하이브리드 전략**은 쌓여있거나 붙어있는 객체를 하나씩 분리하여 파지 가능성을 높이고, 충돌을 줄여 **전체 성공률을 향상**시킵니다. 또 다른 예로, 개방형 세계 그래스핑 (OWG) 연구 (Towards Open-World Grasping with Large Vision-Language Models)에서도 로봇이 바로 파지하지 않고 **"방해되는 주스 상자를 먼저 치운 뒤 목표 장난감을 집는"** 식으로 계획을 세워, 복잡한 씬에서도 높은 완수율을 보였습니다 (Towards Open-World Grasping with Large Vision-Language Models). 이러한 **다중 단계 계획** 접근은 특히 밀집된 객체 더미에서 유효합니다.

- **고성능 인스턴스 세그멘테이션 및 추적**: 복잡한 환경일수록 올바른 **물체 분할과 인식**이 성공적인 파지의 선결조건입니다. Kimhi 등(WACV 2025)은 **상호작용을 통한 학습(Learning-Through-Interaction)**으로 **로봇 인스턴스 세그멘테이션** 성능을 높여, 클러터 속에서 부분 가려져도 **같은 객체를 지속적으로 인식**하도록 했습니다 (Robot Instance Segmentation with Few Annotations for Grasping). 라벨이 적게 달린 데이터로 학습하면서도, **시각 일관성 유지(temporal consistency)** 기법을 도입해 연속 프레임 상에서 객체 마스크의 연속성을 보장함으로써, 가려졌다 나타나도 동일 객체로 인식하는 강인함을 달성했습니다 (Robot Instance Segmentation with Few Annotations for Grasping). 이처럼 **마스크 연속성**과 **정확한 객체 분할**은 잡는 도중 객체를 놓치지 않도록 하며, 복잡한 씬에서 잘못된 물체를 집는 오류를 줄입니다. 더욱이, SAM과 같은 강력한 분할 모델을 **로봇 시각 모듈에 통합**하는 시도 (Language-driven Grasp Detection with Mask-guided Attention)로 가려진 물체의 경계까지 잘 분할해내거나, **Fusion 방식**으로 멀티뷰 관찰을 결합해 보이지 않던 면의 grasp까지 계획하는 연구도 나타났습니다.

- **자연어 지시와 세그멘테이션 활용**: 복잡한 환경에서는 인간 작업자의 도움을 받아 **특정 물체**를 지시하는 경우가 많습니다. 이에 따라 **Referring Grasping** 분야가 발전하여, **자연어로 언급된 객체**를 정확히 집기 위한 기술들이 개발되었습니다 (Language-driven Grasp Detection with Mask-guided Attention). 이는 세그멘테이션으로 장면의 객체들을 분리하고, 언어 입력에 따라 **해당 마스크에 가중치를 주어 파지 계획**을 수립합니다. 예컨대 Vo 등(IROS 2024)의 방법은 **"오른쪽에 있는 빨간 컵 집어"** 같은 명령에 대해, 세그멘테이션 마스크들을 얻은 후 언어-시각 어텐션으로 그 중 타겟 마스크를 선택하고 그리퍼 자세를 산출합니다 (Language-driven Grasp Detection with Mask-guided Attention). 언어 조건이 추가됨으로써 로봇은 복잡한 씬에서도 **사용자가 원하는 정확한 대상**을 파지할 수 있게 되고, 부정확한 인식으로 엉뚱한 물체를 드는 실수를 줄였습니다. 이러한 기술은 클러터된 가정집 환경 등에서 유용하며, 여러 객체 중 **우선순위 대상을 식별**하여 성공률을 높입니다.

- **평가 및 런타임 조정**: 복잡한 환경에서는 한 번의 시도로 실패할 수 있기 때문에, **실시간 평가와 대응**이 중요합니다. 최신 시스템들은 **파지 실패를 감지**하면 즉각 물체를 내려놓고 다른 자세를 시도하거나, **물체가 미끄러질 경우 재조정**하는 피드백 루프를 갖추고 있습니다. 또한 파지 후 들어올릴 때 주변 물체와 부딪치지는 않는지 **경로 재계획**을 수행하여 성공률뿐 아니라 **성공 후 안정적 운반**까지 고려합니다. 예컨대 Xu 등(ICRA 2023) 시스템은 잘못 집혔거나 잡는 중 위치가 어긋난 경우 **한번에 해결하지 않고 추가 동작**(예: 다시 잡기)을 통해 최종 성공을 도모했고, 그 결과 동작 횟수는 늘어날지언정 **최종 작업 완수율**을 높였습니다 (GitHub - xukechun/Vision-Language-Grasping: [ICRA 2023] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter).

이렇듯, 복잡한 환경에서의 그래스핑 성공률을 높이기 위해 **다단계 계획(pushing+grasping)**, **향상된 세그멘테이션 및 추적**, **자연어를 통한 정확한 타겟 지정**, **실시간 피드백 보정** 등의 총체적인 방법론이 연구되고 있습니다. 이를 통해 로봇은 인간 수준의 유연성과 정확성으로 어수선한 환경에서도 목표물을 집어낼 수 있게 발전하고 있습니다.

5. 세그멘테이션 정보를 활용한 그래스핑 구현 기법

세그멘테이션된 **객체 마스크 정보**는 로봇 파지 구현 과정에서 다양하게 활용될 수 있습니다. 주요 활용으로는 **타겟 객체 분리와 식별**, **파지 후보 제한 및 검증**, **그리퍼 파지폭(gripper width) 결정**, **물체 중심 계산 및 접근 방향 보정**, **마스크 연속성 체크** 등이 있습니다. 2023년 이후 연구에서는 세그멘테이션 출력물을 활용해 파지 동작의 신뢰성과 정확성을 높이는 구현 기법들이 제안되었습니다.

- **배경제거 및 관심영역 한정**: 세그멘테이션의 가장 1차적 활용은 **복잡한 배경으로부터 목표 객체 픽셀만 분리**하는 것입니다. Mask-GD 등 기존 연구 (□) (□)에서도 **객체 마스크(MASK)**만을 입력으로 사용하는 그립 검출을 통해, 전체 이미지에서 불필요한 배경 피처로 인한 혼선을 줄이고 **검출 연산 범위를 축소**함으로써 효율과 정확도를 높였습니다. 이는 특히 여러 물체가 있는 장면에서 효과적이며, 세그멘테이션 단계에서 얻은 **ROI(관심 영역)**만을 대상으로 그립 후보를 찾기 때문에 **연산 자원도 절약**됩니다. 최신 비전-언어 파지 연구들도 SAM 등을 이용해 **장면 분할→타겟 마스크 선택** 과정을 선행함으로써, 파지 모듈이 정확히 타겟 객체에만 집중하도록 합니다 ([2302.12610] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter) ([2302.12610] A Joint Modeling of Vision-Language-Action for Target-oriented Grasping in Clutter).

- **파지 후보 필터링 및 마스크 연속성**: 세그멘테이션 정보는 생성된 파지 후보들의 **유효성 검증**에 활용됩니다. 예를 들어 파지 후보가 **타겟 마스크 영역 내에 실제로 들어오는지**를 체크하여, 분할된 객체를 벗어나는 그립은 폐기합니다. 또한 마스크가 **불연속적인 여러 조각**으로 나뉘었다면(과분할 문제), 이를 하나의 객체로 **연결되었는지(continuity)** 확인하는 후처리를 거쳐야 합니다. Tziafas 등(2024)의 OWG 파이프라인에서도 기본 SAM이 **과분할**하는 경향이 있어, **객체 단위로 마스크를 통합**하도록 처리한 바 있습니다 (예: 인접 마스크 병합 또는 GPT-4 Vision에 numeric ID로 표시하여 동일 객체로 묶음) (Academic Project Page). 마스크 연속성 체크를 통해 **그리퍼가 하나의 온전한 물체만 집도록** 보장하며, 둘 이상의 객체를 함께 집어 실패하는 사태를 예방합니다.

- **그리퍼 폭 및 자세 최적화**: 세그멘테이션 마스크로부터 **물체의 크기와 형상**을 추출하면, **그리퍼의 개폐폭**을 적절히 설정할 수 있습니다. 예컨대 사각형 바운딩박스나 마스크의 최대 거리 등을 이용해, 물체를 잡는 데 필요한 최소~최대 간격을 추정하고 그리퍼를 맞춰 엽니다. 실제 많은 그립 검출 NN들은 출력으로 **그립 품질(중심)**, **그립 각도**, **그리퍼 폭**을 예측하는데 (Show and Grasp: Few-shot Semantic Segmentation for Robot Grasping through Zero-shot Foundation Models), 세그멘테이션 기반 방법에서는 **타겟 객체의 분할영역을 히트맵으로 제공**하여 네트워크가 해당 객체의 폭과 방향에 맞는 그립을 내도록 합니다 (Show and Grasp: Few-shot Semantic Segmentation for Robot Grasping through Zero-shot Foundation Models). 이를 통해 **그리퍼 너비를 객체에 최적화**하여 너무 벌리거나 너무 좁게 잡아서 미끄

러지는 일을 막습니다. 또한 마스크의 모양으로 물체의 주된 **방향성**(principal axis)를 파악해 그리퍼를 길이방향으로 놓거나, **긴 막대형 물체는 중앙이 아닌 약간 한쪽을 잡아 안정화**하는 등, 세그멘테이션을 활용한 **자세 보정** 기법도 사용됩니다.

- **객체 중심점 및 무게중심 계산**: 분할된 객체의 **픽셀 분포**와 Depth를 결합하면, 물체의 3차원 **무게중심 추정**이 가능합니다. 로봇은 이를 참고하여 **접근 경로**를 계획하거나 들었을 때 기울어지지 않도록 **중심을 잡는 위치**를 선택합니다. 예를 들어 마스크의 centroid를 계산하고, Depth 정보를 평균내어 3D centroid를 얻은 뒤 그 부근에 파지를 시도하면 보다 안정적인 집기가 됩니다. 특히 **비정형 물체**의 경우 무게중심이 기하학적 중심과 다를 수 있는데, 세그멘테이션된 점군을 이용해 무게 분포를 추정하는 기법이 유용합니다. 일부 연구에서는 파지 전후의 **마스크 변화**를 통해 객체가 그리퍼에 잘 들어왔는지(예: 들어올렸을 때 마스크가 사라지면 성공) 판단하기도 합니다. 이러한 **마스크 기반 모니터링**은 파지 동작의 신뢰도를 높여주는 구현상의 트릭입니다.

- **실시간 피드백 연계**: 세그멘테이션은 로봇 제어 피드백에도 활용됩니다. 예를 들어 파지하는 동안 연속 프레임에서 목표 마스크를 추적하여, 만약 그립 과정에서 **마스크가 급격히 움직이거나 이탈**하면 물체가 떨어졌음을 감지합니다. 또는 로봇 팔이 접근할 때 **마스크 크기가 점점 줄어들면** 카메라에 가려지고 있음을 의미하므로, 이 정보를 토대로 속도를 조절하거나 카메라 시점을 바꾸는 등의 대응을 할 수 있습니다. 최신 연구에서는 이러한 **세그멘테이션 피드백 루프**를 통해 더욱 **견고한 파지 동작**을 구현하고 있습니다.

정리하면, 세그멘테이션 정보는 로봇 그래스핑 파이프라인의 여러 단계에서 **환경에 대한 명확한 인식과 수치적 피드백**을 제공합니다. 이를 통해 파지 동작을 더 정밀하게 제어하고 실패를 줄이는 다양한 구현 기법이 연구 및 적용되고 있습니다 (Show and Grasp: Few-shot Semantic Segmentation for Robot Grasping through Zero-shot Foundation Models) (Robot Instance Segmentation with Few Annotations for Grasping). 실제 오픈소스 로봇 시스템에서도 이러한 아이디어(예: 마스크로 객체 크기 추정해 그리퍼 제어)를 적극 도입하는 추세입니다.

6. Jetson AGX Orin 실시간 시스템을 위한 최적화 전략

Jetson AGX Orin과 같은 엣지 AI 디바이스에서 **그래스핑 파이프라인의 실시간성**을 보장하려면, 한정된 **계산 자원(CUDA 코어, DLA, 메모리)**을 효율적으로 활용하는 최적화가 필수입니다. 2023년 이후 발표된 시스템들은 **모델 경량화**, **하드웨어 가속기 활용**, **병렬 파이프라인**, **모델 통합** 등 다양한 기법으로 Orin 상의 **엔드투엔드 지연 시간**을 줄이고 있습니다.

- **모델 경량화와 TensorRT 최적화**: NVIDIA Jetson AI Lab에서는 대형 비전 모델들을 Orin에 맞게 경량화하고 TensorRT로 최적화한 **nano** 시리즈를 공개했습니다. 앞서 언급한 **nanoOWL**(OWL-ViT 최적화)과 **nanoSAM**(SAM 경량화)이 대표적이며, 이들은 FP16이나 INT8 정밀도로 변환하고 불필요한 연산을 줄여 **Jetson Orin에서 수십 FPS 실시간** 동작을 달성했습니다 (GitHub - NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT.) (NVIDIA-AI-IOT/nanosam: A distilled Segment Anything (SAM) model ...). **NanoSAM**의 경우 **지식 증류(distillation)**를 통해 원본 SAM의 성능을 유지하면서도 모델 크기를 크게 줄였으며, **MobileSAM**이나 **EfficientSAM** 등의 연구에서도 **압축 백본**이나 **가지치기+양자화(pruning+quantization)**로 **연산량을 최소화**하는 기법을 도입했습니다 (GraspSAM: When Segment Anything Model Meets Grasp Detection). 이러한 경량 모델을 TensorRT로 엔진화하면, Orin의 GPU와 DLA에서 최적 스케줄로 실행되어 레이턴시를 크게 단축할 수 있습니다. 예를 들어 nanoOWL은 Jetson Orin Nano에서도 실시간(약 30FPS) 성능을 보였고, AGX Orin에서는 그 이상을 낼 수 있다고 보고되었습니다 (GitHub - NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT.).

- **파이프라인 병렬화 및 자원 분배**: 실시간 시스템에서는 인식->계획->제어의 파이프라인을 **동시병렬적으로 처리**하여 효율을 높입니다. Orin은 멀티코어 CPU와 GPU가 있으므로, 예컨대 **한 프레임에서 세그멘

테이션 계산을 GPU에서 하는 동안 CPU에서는 **이전 프레임의 파지 경로를 계획**하도록 파이프라인을 구성합니다. 또한 여러 DNN 모델(검출, 분할, 포즈산출 등)을 순차 대신 병렬 실행하거나, 필요에 따라 한 모델의 출력을 다음 모델이 기다리지 않고 스트리밍 처리하도록 합니다. 일부 연구에서는 **Asynchronous pipeline** 설계를 통해 **센서 입력부터 그리퍼 제어 명령까지 지연을 최소화**했으며, 이것이 실제 로보컵 등 경쟁에서 성능 향상으로 이어졌습니다. Jetson Orin의 경우 하나의 대형 GPU보다는 **작은 CUDA 스트림들의 동시 실행 최적화**에 강점이 있으므로, 이를 활용하도록 소프트웨어 스레드와 CUDA 커널을 튜닝하는 것이 중요합니다.

- **단일 모델 다기능화**: 복수의 모델을 사용하면 각 모델의 실행 지연이 누적되고 메모리 사용도 증가합니다. 이를 개선하고자, **하나의 모델이 여러 기능을 수행하도록 통합**하는 접근이 있습니다. 예를 들어 **GraspSAM**은 물체 세그멘테이션과 그립 검출을 하나의 네트워크로 합쳤기 때문에, 별도의 검출+파지 두 모델을 사용할 때보다 **계산량과 메모리 요구가 감소**했습니다 (GraspSAM: When Segment Anything Model Meets Grasp Detection) (GraspSAM: When Segment Anything Model Meets Grasp Detection). 이러한 통합 모델은 **엔드투엔드 추론 시간**을 크게 단축시켜 실시간성에 유리합니다. 다만 복합 모델이 너무 커지면 오히려 불리하므로, GraspSAM처럼 **최소한의 토큰 학습과 약간의 파라미터 추가**만으로 통합하는 **경량 통합 전략**이 연구되고 있습니다 (GraspSAM: When Segment Anything Model Meets Grasp Detection).

- **저전력 모드와 발열 관리**: Jetson AGX Orin은 강력하지만 발열과 전력제한이 있어 **지속적 실시간 구동** 시 클럭 스케일링이 일어날 수 있습니다. 이를 방지하기 위해 **전력 모드 설정**(MAXN 등)과 적극적 쿨링, 그리고 **연산 부하 균형화**가 필요합니다. 예컨대 파지 경로 계산과 같은 비교적 가벼운 연산은 CPU에서 처리하여 GPU 연산이 쉬는 시간을 주거나, 연산량이 많은 프레임은 간헐적으로 drop하여 평균 부하를 조절하는 방법도 있습니다. 실제 현업에서는 Orin에서 **50~60FPS 이상의 카메라 스트림**을 처리할 때 **프레임 스킵** 기술과 **동적 해상도 조절**을 통해 부하를 관리하며 실시간성을 유지합니다.

- **I/O 및 기타 최적화**: 실시간 로봇 시스템에서는 **카메라로부터 영상 입수, 그리고 제어 신호 출력**까지의 모든 단계가 최적화 대상입니다. 메모리 복사 횟수를 줄이기 위해 **Zero-copy DMA**로 카메라 프레임을 GPU 메모리에 바로 올리거나, TensorRT에 맞게 **NHWC->NCHW 전치** 등을 사전에 해두는 등이 활용됩니다. 또한 Jetson의 ISP나 PVA 같은 모듈을 활용해 **전처리**(예: 리사이즈, 컬러공간 변환)를 CPU 개입 없이 처리하기도 합니다. 이렇듯 하드웨어의 모든 기능을 활용한 최적화로 **엔드투엔드 지연 수십 ms 수준**의 그래스핑 시스템 구현 사례도 보고되고 있습니다.

요약하면, Jetson AGX Orin 상에서 그래스핑 시스템을 실시간 구동하려면 **모델 경량화 및 가속(예: NanoSAM, NanoOWL)** (GitHub - NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT.) (NVIDIA-AI-IOT/nanosam: A distilled Segment Anything (SAM) model ...), **병렬 파이프라인 설계**, **모델 통합을 통한 연산 절약** (GraspSAM: When Segment Anything Model Meets Grasp Detection), 그리고 **하드웨어 리소스 세부 튜닝**이 요구됩니다. 이러한 최적화 전략을 통해, 엣지 디바이스에서도 복잡한 비전-로보틱스 파이프라인이 끊김 없이 동작하여 즉각적인 로봇 대응이 가능해지고 있습니다. 최근 대부분의 관련 논문들은 자체 코드를 공개하여, 실무 개발자들도 해당 최적화 기법과 모델을 활용할 수 있도록 지원하고 있습니다.