최신 컴퓨터 비전 모델 및 기술 보고서

1. 서론

컴퓨터 비전 분야는 이미지 분할 및 객체 인식과 같은 영역에서 괄목할 만한 발전을 거듭해 왔습니다. 연구자와 실무자 모두에게 다양한 모델과 기술의 미묘한 차이를 이해하는 것은 매우 중요합니다. 이 보고서는 "SAM2", "ZegCLIP", "SCLIP", "CLIP-SEG", "NanoOWL", "Semantic-SAM", "Semantic-Segment-Anything", "EVLA", "CLIP-RT", "NanoSAM", "grounding dino", "grounded-sam-2", "openvla", "MobileSAMv2", "ESAM"의 15가지 특정 컴퓨터 비전 기술의 개념, 기능, 응용 분야 및 성능에 대한 자세한 분석을 제공하는 것을 목표로 합니다.

2. 개별 기술 심층 분석

2.1. SAM2 (Segment Anything Model 2)

- 개념적 개요 및 목표: Meta의 차세대 비디오 및 이미지 분할 모델인 SAM2는 기존 SAM을 기반으로 합니다.[1] 주요 목표는 이미지와 비디오에서 모든 객체를 빠르고 정확하게 분할하는 것입니다.[4] SAM2는 특정 이미지에 대해 학습할 필요 없이 제로샷 분할을 수행하여 학습 중에 보지 못한 객체를 처리할 수 있습니다.[3] 클릭, 상자 및 마스크와 같은 다양한 프롬프트를 사용할 수 있으며 [2], 이미지 및 비디오 작업 모두에 통합된 아키텍처를 활용합니다.[1] 비디오로의 확장과 아키텍처의 통합은 기존 SAM에 비해 상당한 발전으로, 컴퓨터 비전에서 보다 다재다능한 기본 모델로 나아가고 있음을 시사합니다. 이미지와 비디오를 단일 모델로 처리할 수 있는 통합은 개발 및 배포를 단순화하여 두 가지 기능 모두가 필요한 응용 프로그램에 유용합니다. 기존 SAM과의 비교는 명확한 발전과 기능 확장을 보여줍니다.
- 핵심 기능 및 작동 원리: SAM2는 대상 객체에 대한 문맥 정보를 유지하면서 비디오 프레임을 순차적으로 처리하도록 설계된 메모리 장착 트랜스포머 아키텍처를 사용합니다.[1] 비디오 전체에서 객체의 공간적 범위를 정의하기 위해 개별 프레임에 포인트, 상자 및 마스크 프롬프트를 사용할 수 있습니다.[2] 스트리밍 아키텍처는 긴 비디오와 로봇 공학과 같은 실시간 응용 프로그램에서 효율적인 처리를 가능하게 합니다.[2] 비디오 분할을 위해 메모리 인코더, 메모리 뱅크 및 메모리 주의 모듈이 프레임 정보와 사용자 상호 작용을 저장하는 데 사용됩니다.[2] 후속 프레임에 프롬프트를 추가하여 반복적인 개선 프로세스가 가능합니다.[2] 프롬프트가 모호할 때 여러 마스크를 생성할 수 있습니다.[2] 이미지 분할의 경우 경량의 프롬프트 가능한 마스크 디코더를 사용하여 SAM과 유사하게 작동합니다.[2] 메모리 메커니즘의 도입은 비디오 처리에 매우 중요하며, 프레임 간의 컨텍스트를 유지하여 이미지 전용 모델과 차별화됩니다. 메모리 구성 요소(인코더, 뱅크, 주의)와 스트리밍 아키텍처에 대한 자세한 설명은 SAM2가 비디오의 시간적 차원을 처리하는 방법을 명확히 보여줍니다. 이는 단순히 각 프레임에 독립적으로 이미지 분할 모델을 적용하는 것을 넘어섭니다.

- 고유한 특징 및 변형: SAM2는 초당 약 44프레임의 실시간 성능을 제공합니다.[2] 이전 접근 방식에 비해 효율성이 향상되었고 필요한 사용자 상호 작용이 줄었습니다.[1] 빠른 모델과 고해상도 모델과 같은 특수 배포 모델을 사용할 수 있습니다.[1] 또한 폐색 처리 기능도 강조할 만한 특징입니다.[3] 실시간 성능은 대화형 응용 프로그램에 SAM2를 적합하게 만들고, 필요한 상호 작용 감소는 사용자 경험을 향상시킵니다. 다양한 배포 모델은 속도와 정확성에 대한 다양한 응용 프로그램 요구 사항을 충족합니다.
- 실제 응용 분야 및 사용 사례: SAM2는 비디오 편집, 증강 현실, 감시, 스포츠 분석, 환경 모니터링, 전자 상거래 및 자율 주행 차량과 같은 분야에서 잠재적인 응용 분야를 가지고 있습니다.[3] 데이터 주석 및 객체 추적에도 사용할 수 있습니다.[6] 비디오 생성 모델에 대한 입력으로 사용하여 정확한 편집 기능을 활성화할 수 있는 잠재력도 있습니다.[4] 광범위한 응용 분야는 시각 분석과 관련된 다양한 산업 및 작업에서 SAM2의 다재다능함을 강조합니다. 다양한 응용 프로그램 예는 이미지와 비디오 모두에서 작동하는 기본 분할 모델의 광범위한 적용 가능성을 보여줍니다.
- 성능 지표 및 평가: SAM2는 이전 접근 방식에 비해 3배 적은 사용자 상호 작용으로 뛰어난 정확도를 제공합니다.[1] SA-23 벤치마크에서 58.9%의 1-클릭 mIoU를 달성하여 SAM의 58.1%를 개선했습니다.[1] 이미지 분할을 위해 원래 SAM에 비해 6배 빠른 추론 속도를 보입니다.[1] 또한 17개의 제로샷 비디오 데이터 세트에서 이전 방법보다 훨씬 뛰어난 결과를 달성했으며, 약 3배 적은 사용자 상호 작용이 필요했습니다.[2] 제로샷 벤치마크 스위트에서 SAM보다 6배 빠르며 DAVIS, MOSE, LVOS 및 YouTube-VOS와 같은 확립된 비디오 객체 분할 벤치마크에서 뛰어납니다.[2] 정량적 성능 지표(mIoU, 속도 향상)는 정확성과 효율성 모두에서 이전 모델 및 기타 기존 방법에 비해 SAM2의 발전을 명확하게 보여줍니다. 특정 벤치마크 결과와 SAM 및 기타 비디오 분할 모델과의 비교는 주장된 우수한 성능에 대한 증거를 제공합니다.

2.2. ZegCLIP

- 개념적 개요 및 목표: ZegCLIP은 제로샷 시맨틱 분할을 위해 CLIP을 적용하는 단일 단계 방법으로 소개되었습니다.[7] 주요 목표는 일반화된 제로샷 시맨틱 분할을 위해 시각적 이해와 언어적 이해 사이의 간극을 메우는 것입니다.[8] ZegCLIP은 CLIP에서 경량 디코더로 직접 지식을 전이합니다.[7] ZegCLIP은 작업별 학습 데이터 없이 분할을 수행하기 위해 강력한 사전 학습된 CLIP의 지식을 활용하여 시맨틱 분할을 위한 비전-언어 모델의 잠재력을 보여줍니다.
- 핵심 기능 및 작동 원리: VLM(Vision-Language Model)의 시각적 인코더는 분할 모델의 인코더 역할을 하며 [8], 정렬된 텍스트 인코더는 각 클래스 레이블 이름에 대한 클래스 임베딩을 생성합니다.[8] RPM(Relationship Prompt Module)은 이미지 및 텍스트 인코딩 레이어의 출력을 활용하여 픽셀 수준 관계 프롬프팅을 가능하게 합니다.[7] VLM에서 계층별 안내 모드를 구현하여 이미지에서 픽셀 수준으로 임베딩을 점진적으로 전송할 수 있습니다.[7] ZegFormer가 객체 토큰에서 교차 모달리티 정렬을

수행하는 것과 달리 ZegCLIP은 패치 토큰에서 텍스트-패치 매칭을 구현합니다.[8] ZegCLIP의 핵심은 CLIP의 시각적 및 텍스트 인코딩 기능을 미세한 수준에서 분할 프로세스를 안내하는 데 효과적으로 활용하는 데 있습니다.

- 고유한 특징 및 변형: Zsseg와 같은 2단계 방법에 비해 단일 단계 방식으로 작동합니다.[7] 상당한 계산 오버헤드 없이 정확한 특징 지역성과 강력한 텍스트 정렬을 달성합니다.[8] 낮은 계산 오버헤드는 ZegCLIP을 계산 집약적인 모델에 비해 실제 응용 프로그램에 더 실용적으로 만들 수 있는 귀중한 기능입니다.
- 실제 응용 분야 및 사용 사례: 스니펫에서 명시적으로 언급되지는 않았지만 제로샷 기능은 전문화된 산업 검사 또는 과학 연구와 같이 새롭거나 보이지 않는 객체 범주가 있는 시나리오에서 응용 분야를 시사합니다. 투명 객체(안경)와 같이 어려운 경우에 향상된 분할 결과를 위해 SAM2와 통합할 수 있는 기능은 특정 도메인에서 향상된 성능을 위한 잠재력을 나타냅니다.[8]
- 성능 지표 및 평가: MFuser를 ZegCLIP에 적용하여 합성에서 실제 벤치마크에서 68.20 mIoU, 실제에서 실제 벤치마크에서 71.87 mIoU를 달성했습니다.[8] 보고된 mIoU 점수는 시맨틱 분할 작업에서 ZegCLIP의 성능에 대한 정량적 측정값을 제공합니다.

2.3. SCLIP (Semi-supervised CLIP)

- 개념적 개요 및 목표: S-CLIP은 준지도 학습 방식을 사용하여 CLIP에서 순진한 의사 레이블링 문제를 해결하기 위해 도입되었습니다.[9] 주요 동기는 이미지-텍스트 쌍이 제한적인 시나리오에서 쌍을 이루지 않은 이미지를 통합하여 CLIP 학습을 개선하는 것입니다.[9] S-CLIP은 레이블이 지정된 데이터가 부족한 비전-언어 작업에서 레이블이 지정되지 않은 풍부한 데이터를 효과적으로 활용하여 모델 성능을 향상시키는 것을 목표로 합니다.
- 핵심 기능 및 작동 원리: S-CLIP은 캡션 수준 및 키워드 수준 의사 레이블링 접근 방식을 도입합니다.[9] 이는 open_clip 라이브러리를 기반으로 구현됩니다.[9] 제안된 학습 손실의 주요 논리는 custom/loss.py에 구현되어 있습니다.[9] 캡션 수준 및 키워드 수준 의사 레이블링을 모두 사용하면 쌍을 이루지 않은 이미지에서 다양한 수준의 의미 정보를 추출하는 전략을 시사합니다.
- 고유한 특징 및 변형: 완전 지도 학습 방식인 CLIP과 달리 준지도 학습 방식이라는 점이 특징입니다.[9] 제한된 전문 캡션으로 성능을 향상시키는 데 중점을 둡니다.[9] S-CLIP은 고품질의 쌍을 이룬 이미지-텍스트 데이터의 대규모 데이터 세트를 얻는 것이 어려운 실제 시나리오에서 실용적인 이점을 제공합니다.
- 실제 응용 분야 및 사용 사례: 전문 의료 영상 또는 틈새 제품 카탈로그와 같이 쌍을 이룬 이미지-텍스트 데이터를 수집하는 데 비용이 많이 들거나 시간이 많이 걸리는 도메인에 적합합니다. S-CLIP은 데이터 제한적인 도메인에서 CLIP과 유사한 모델을 적용하기 위한 진입 장벽을 낮출 수 있습니다.
- 성능 지표 및 평가: 스니펫은 구현 및 동기에 대한 정보를 제공하지만 특정 성능 지표는 부족합니다. 낮은 데이터 체제에서 표준 CLIP에 대한 성능 향상을 정량화하려면 추가 연구가 필요합니다. 개념은 유망하지만 S-CLIP의 효과에 대한

경험적 평가는 매우 중요합니다.

2.4. CLIP-SEG

- 개념적 개요 및 목표: CLIPSeg는 제로샷 및 원샷 이미지 분할을 위해 고정된 CLIP 모델 위에 최소한의 디코더를 추가하는 모델로 소개되었습니다.[10] 주요 목표는 텍스트 및 이미지 프롬프트를 포함하여 테스트 시간에 임의의 프롬프트에 따라 이미지 분할을 생성하는 것입니다.[10] 또한 참조 표현 분할, 제로샷 분할 및 원샷 분할을 통합 모델로 처리할 수 있습니다.[10] CLIPSeg는 CLIP의 학습된 표현의 뛰어난 적응성을 보여줍니다.
- 핵심 기능 및 작동 원리: 아키텍처는 고정된 CLIP 모델과 밀집 예측을 용이하게 하는 트랜스포머 기반 디코더로 구성됩니다.[11] 텍스트(input_ids) 또는 이미지(conditional_pixel_values)를 프롬프트로 사용할 수 있습니다.[10] PhraseCut 데이터 세트의 확장된 버전으로 학습하여 이진 분할 맵 생성을 향상시킵니다.[10] CLIP 백본에 디코더를 추가하면 모델이 분할 마스크를 출력할 수 있습니다.
- 고유한 특징 및 변형: 재학습 없이 다양한 분할 작업에 동적으로 적응할 수 있다는 점이 특징입니다.[11] 텍스트 및 이미지 프롬프트를 모두 활용하는 하이브리드 입력 메커니즘을 강조합니다.[11] 또한 새로운 분할 작업에 빠르게 적응할 수 있는 다재다능함과 효율성을 제공합니다.[12] 이미지 프롬프트를 분할에 사용할 수 있는 기능은 강력한 기능입니다.
- 실제 응용 분야 및 사용 사례: 의료 영상(새로운 해부학적 구조 분할), 자율 주행 차량(예상치 못한 장애물 식별) 및 농업(식물 질병 감지) 분야에서 응용 분야가 언급되었습니다.[12] 또한 증강 현실에서 실제 객체를 분할하는 데에도 사용할 수 있습니다.[13] 제로샷 특성으로 인해 새로운 객체 범주가 자주 나타나는 도메인에서 특히 유용합니다.
- 성능 지표 및 평가: 스니펫은 특정 정량적 성능 지표보다는 기능 및 아키텍처에 더 중점을 둡니다. 자세한 평가 결과는 원본 논문(Lüddecke 및 Ecker)에 포함되어 있을 가능성이 높습니다.

2.5. NanoOWL

- 개념적 개요 및 목표: NanoOWL은 NVIDIA TensorRT를 사용하여 NVIDIA Jetson Orin 플랫폼에서 실시간 추론을 위해 OWL-ViT를 최적화하는 프로젝트로 소개되었습니다.[14] 주요 목표는 에지 장치에서 빠르고 효율적인 개방형 어휘 객체 감지를 활성화하는 것입니다.[14] 또한 텍스트 프롬프트를 제공하여 모든 수준에서 중첩된 감지 및 분류를 가능하게 하는 OWL-ViT와 CLIP을 결합한 "트리 감지" 파이프라인을 도입합니다.[14] NanoOWL은 리소스가 제한된 장치에서 효율적인 실시간 개방형 어휘 객체 감지의 필요성을 해결합니다.
- 핵심 기능 및 작동 원리: OWL-ViT(Vision Transformers를 사용한 개방형 세계 객체 감지)를 기반으로 합니다.[14] 트리 감지 파이프라인은 텍스트 프롬프트를 제공하여 모든 수준에서 객체를 감지하고 분류할 수 있습니다.[14] 트리 예측 파이프라인에서

감지를 위해 OWL-ViT를, 분류를 위해 CLIP을 결합합니다.[14] NVIDIA TensorRT를 사용하여 추론을 최적화합니다.[14] "트리 감지" 메커니즘은 이미지에서 객체를 계층적으로 이해하고 분류하는 방법을 제공합니다.

- 고유한 특징 및 변형: Jetson Orin Nano에서 실시간 성능을 제공합니다.[14] 에지 배포에 최적화되어 있습니다.[16] 중첩된 감지 및 분류를 수행할 수 있습니다.[14] 실시간 성능과 에지 배포에 중점을 두는 것은 로봇 공학, IoT 장치 및 기타 임베디드 시스템에 매우 적합합니다.
- 실제 응용 분야 및 사용 사례: 제로샷 개방형 어휘 인스턴스 분할을 위해 NanoSAM과 결합할 수 있습니다.[14] 로봇 공학에서 객체 감지 및 조작에 사용할 수 있습니다.[15] 지능형 카메라 및 스마트 드론에도 사용할 수 있습니다.[18] NanoSAM과의 잠재적인 시너지 효과는 감지 및 분할을 포함하여 에지 장치에서 포괄적인 장면 이해를 위한 경로를 시사합니다.
- 성능 지표 및 평가: 다양한 OWL-ViT 모델에 대해 Jetson Orin Nano 및 AGX Orin에서 FPS(Frames Per Second)로 성능 수치를 제공합니다.[14] 다양한 모델에서 달성한 정확도(mAP)를 언급합니다.[14] Jetson Orin Nano에서 "슈퍼 모드"를 활성화했을 때 성능 향상을 언급합니다.[19] 제공된 FPS 및 mAP 값은 특정 하드웨어에서 NanoOWL의 속도와 정확도에 대한 정량적 평가를 제공합니다.

2.6. Semantic-SAM

- 개념적 개요 및 목표: Semantic-SAM은 원하는 세분성으로 모든 것을 분할하고 인식하기 위한 증강된 이미지 분할 기초로 소개되었습니다.[21] 주요 목표는 세분성 제어 가능성과 의미 인식이라는 두 가지 고유한 장점을 제공하는 것입니다.[21] 객체에서 부품까지 원하는 세분성으로 분할 마스크를 생성할 수 있습니다.[21] 또한 의미 레이블을 동시에 예측합니다.[21] Semantic-SAM은 단순한 분할을 넘어 다양한 수준의 의미 이해를 제공하여 보다 포괄적인 장면 해석을 제공합니다.
- 핵심 기능 및 작동 원리: 객체 및 부품 분류를 위한 분리된 접근 방식을 설명하며, 풍부한 분할 데이터 세트 간의 지식 전이를 허용합니다.[21] 객체와 부품을 독립적으로 인코딩하기 위해 공유 텍스트 인코더를 사용합니다.[21] MaskDINO를 따라 쿼리 기반 마스크 디코더를 사용합니다.[21] 일반 쿼리(MaskDINO에서 사용)와 대화형 분할을 위한 공간 쿼리(점 및 상자)의 두 가지 유형의 쿼리를 사용합니다.[21] 다양한 세분성의 마스크를 캡처하기 위해 디코더 아키텍처에 다중 선택 학습 설계가 통합되어 있습니다.[22] 분리된 분류와 공유 텍스트 인코더의 사용은 다양한 수준의 의미 주석이 있는 다양한 분할 데이터 세트를 활용하는 효율적인 방법을 시사합니다.
- 고유한 특징 및 변형: SAM에 비해 세분성 제어 가능성과 의미 인식이라는 고유한 장점을 강조합니다.[21] 일반, 부품 및 클래스에 구애받지 않는 분할 데이터를 포함한 여러 유형의 분할 데이터를 사용할 수 있는 범용 분할 프레임워크를 강조합니다.[21] Semantic-SAM은 의미 이해를 분할 프로세스에 직접 통합하여 이전 모델에 비해 보다 다재다능하고 유익한 분할 모델을 목표로 합니다.
- 실제 응용 분야 및 사용 사례: 세밀한 객체 이해, 상세한 이미지 분석 및 부품 수준

분할이 필요한 작업에 잠재적인 응용 분야가 있습니다. 마스크 영역을 얻기 위해 ComfyUI에서 인페인팅의 사전 노드로 사용됩니다.[23] Semantic-SAM의 기능은 객체의 부품을 식별하는 것만큼 객체 자체를 이해하는 것이 중요한 응용 프로그램에 특히 적합합니다.

- 성능 지표 및 평가: SAM 데이터를 추가하여 COCO 전경 분할에서 60.2 PQ의 새로운 최고 성능을 달성한 것을 언급합니다.[21] 향상된 전경 분할 성능은 Semantic-SAM의 학습 접근 방식의 효과를 입증합니다.

2.7. Semantic-Segment-Anything

- 개념적 개요 및 목표: 이 용어는 Semantic-SAM과 같이 이미지에서 "무엇이든" 분할하고 의미를 이해할 수 있는 분할 모델의 광범위한 목표를 나타내는 것 같습니다.[5] SAM(무엇이든 분할)과 의미 분할(세그먼트가 무엇인지 이해)의 기능을 결합한다는 아이디어를 포함합니다.[24] 목표는 이미지의 모든 영역에 대해 정확한 분할과 의미 있는 레이블링을 모두 달성하는 것입니다. "Semantic-Segment-Anything"은 컴퓨터 비전의 높은 수준의 목표를 나타내며, 시각적 장면의 기하학적 및 의미적 이해를 모두 제공하는 모델 개발을 주도합니다.
- 핵심 기능 및 작동 원리: 개별 객체를 식별하는 인스턴스 분할과 각 픽셀에 클래스 레이블을 할당하는 의미 분할을 모두 수행할 수 있는 모델을 포함합니다.[24] 종종 SAM과 같은 기본 모델을 활용하고 의미 분류 기술과 통합합니다.[21] SAM에서 예측한 마스크에 레이블을 할당하기 위해 CLIP을 사용하는 것과 같은 기술을 포함할 수 있습니다.[28] "Semantic-Segment-Anything"을 달성하려면 분할 및 의미 이해에 특화된 다양한 모델의 강점을 결합하는 모듈식 접근 방식이 필요합니다.
- 고유한 특징 및 변형: 주요 특징은 클래스에 구애받지 않는 분할과 의미 이해의 조합입니다. 이러한 조합을 달성하는 방법에는 다중 작업 학습, 분리된 접근 방식 또는 분류 모델을 사용한 분할 마스크의 사후 처리와 같은 변형이 존재합니다. "Semantic-Segment-Anything"에 대한 다양한 접근 방식은 이러한 두 가지 중요한 장면 이해 측면을 통합하는 가장 효과적인 방법을 찾는 지속적인 연구를 반영합니다.
- 실제 응용 분야 및 사용 사례: 자율 주행(운전 장면의 모든 요소 이해), 의료 영상(다양한 조직 또는 이상 징후 분할 및 식별), 로봇 공학(탐색 및 조작을 위한 상세한 환경 이해) 및 이미지 편집(의미적으로 인식된 이미지 영역 조작)을 포함한 다양한 분야에서 수많은 응용 분야가 있습니다.[24] 장면의 기하학적 구조와 의미를 모두 이해하는 능력은 컴퓨터 비전의 많은 실제 응용 프로그램에 기본적입니다.
- 성능 지표 및 평가: 성능은 분할(예: mIoU, PQ) 및 의미 분류 정확도와 관련된 지표를 사용하여 평가됩니다. 특정 지표는 사용된 특정 모델 및 데이터 세트에 따라 달라집니다. "Semantic-Segment-Anything"을 평가하려면 분할 및 의미 레이블링 모두에서 성능을 고려해야 합니다.

2.8. EVLA (Expanded Very Large Array)

- 개념적 개요 및 목표: 스니펫은 EVLA에 대한 두 가지 뚜렷한 의미를 나타냅니다.

    - 정맥 내 레이저 절제술: 다리 정맥류를 관리하기 위한 최소 침습 의료 치료법입니다.[29] 이는 컴퓨터 비전 모델의 맥락에서 의도된 의미가 아닐 가능성이 높습니다.
    - 확장된 매우 큰 배열: 천문 관측에 사용되는 전파 간섭계입니다.[30] 이것도 목록에 있는 다른 컴퓨터 비전 모델과 직접적인 관련이 없을 가능성이 높습니다. 다른 나열된 항목의 맥락을 고려할 때 EVLA는 이러한 스니펫에 포착되지 않은 보다 최근의 컴퓨터 비전 맥락에서 약어 또는 모델을 나타낼 수 있습니다.
    - 핵심 기능 및 작동 원리: EVLA가 정맥 내 레이저 절제술을 의미하는 경우 레이저 에너지를 사용하여 정맥류를 가열하고 닫는 방식으로 작동합니다.[29] 확장된 매우 큰 배열을 의미하는 경우 여러 전파 망원경을 사용하여 고해상도 전파 하늘 이미징을 위한 큰 구경을 합성하는 방식으로 작동합니다.[30]
    - 고유한 특징 및 변형: 다시 말하지만 이는 EVLA 해석에 따라 달라집니다.
    - 실제 응용 분야 및 사용 사례: 정맥 내 레이저 절제술은 정맥류에 대한 의료 치료에 사용됩니다.[29] 확장된 매우 큰 배열은 다양한 천문 연구 프로젝트에 사용됩니다.[30]
    - 성능 지표 및 평가: 성능 지표는 각 도메인에 따라 다릅니다(예: 의학에서 EVLA의 임상 성공률, 천문학에서 EVLA의 이미지 감도 및 해상도).

2.9. CLIP-RT (CLIP-based Robotics Transformer)

- 개념적 개요 및 목표: CLIP-RT는 일반적인 조작 정책을 위한 비전-언어-액션(VLA) 모델로 소개되었으며, OpenAI의 CLIP을 로봇 학습으로 원활하게 확장합니다.[32] 주요 목표는 이미지와 언어 명령이 주어졌을 때 자연어로 지정된 로봇 동작을 예측하는 방법을 배우는 것입니다.[32] CLIP-RT는 시각, 언어 및 로봇 동작 사이의 간극을 메워 로봇이 시각적 입력과 자연어 명령에 따라 지침을 이해하고 실행할 수 있도록 합니다.
- 핵심 기능 및 작동 원리: 사전 학습된 CLIP 모델을 로봇 학습에 적용하여 대조적 모방 학습을 통해 언어 기반 동작 기본 요소를 학습합니다.[33] 이미지와 지침에 따라 자연어로 지정된 로봇 동작을 예측할 수 있습니다.[32] 예측된 언어 동작을 저수준 로봇 동작으로 변환하기 위해 조회 테이블을 사용합니다.[32] Open X-Embodiment 데이터 세트에서 사전 학습됩니다.[32] 언어 기반 동작 기본 요소의 사용은 모델이 로봇 동작의 보다 추상적이고 잠재적으로 더 일반화 가능한 표현을 학습할 수 있도록 합니다.
- 고유한 특징 및 변형: 새로운 조작 작업을 위한 엔드 투 엔드 로봇 정책을 효과적으로 학습할 수 있다는 점이 특징입니다.[32] 대조적 모방 학습 접근 방식을 강조합니다.[33] 매개 변수가 7배 적은 상태에서 최첨단 OpenVLA 모델보다 성능이 뛰어납니다.[33] 적은 매개 변수로 상당한 성능 향상을 보이는 것은 CLIP-RT 접근 방식의 효율성과 효과를 강조합니다.
- 실제 응용 분야 및 사용 사례: 다양한 환경에서 로봇을 위한 일반적인 조작 정책.[32] 프레임워크에서 수집한 도메인 내 데이터로 미세 조정하여 다양한 로봇 기술 학습.[33] CLIP-RT는 로봇이 더 넓은 범위의 작업을 더 큰 자율성과 적응성으로 수행할 수

있도록 하는 잠재력을 가지고 있습니다.

- 성능 지표 및 평가: OpenVLA(7B 매개 변수)보다 평균 성공률이 24% 높으면서 매개 변수가 7배 적습니다(1B).[33] 소량의 데이터로 일반화하는 능력에서 상당한 개선을 보입니다.[33] 적은 매개 변수로 상당한 성능 향상을 보이는 것은 CLIP-RT 접근 방식의 효율성과 효과를 강조합니다.

2.10. NanoSAM

- 개념적 개요 및 목표: NanoSAM은 NVIDIA TensorRT를 사용하여 NVIDIA Jetson Orin 플랫폼에서 실시간 성능을 위해 설계된 Segment Anything Model(SAM)의 증류된 변형으로 소개되었습니다.[14] 주요 목표는 에지 배포에 적합한 가속화된 모든 것 분할 기능을 달성하는 것입니다.[17] MobileSAM 이미지 인코더를 레이블이 지정되지 않은 이미지에 증류하여 학습됩니다.[17] NanoSAM은 SAM의 강력한 분할 기능을 효율성을 크게 향상시켜 리소스가 제한된 에지 장치에서 사용할 수 있도록 합니다.
- 핵심 기능 및 작동 원리: 파이프라인은 원래 SAM과 유사하지만 경량 이미지 인코더(ResNet18)와 MobileSAM 마스크 디코더를 사용합니다.[17] NVIDIA TensorRT를 사용하여 실시간으로 작동하여 추론을 최적화합니다.[14] 경계 상자 및 키포인트와 같은 다양한 프롬프트 방법을 지원합니다.[17] 증류 프로세스를 통해 NanoSAM은 SAM의 많은 기능을 유지하면서 크기가 훨씬 작고 속도가 빨라집니다.
- 고유한 특징 및 변형: Jetson Orin Nano에서 실시간 성능을 제공합니다.[14] 원래 SAM 및 MobileSAM에 비해 모델 크기가 훨씬 작고 추론 속도가 빠릅니다.[17] 원래 SAM 파이프라인과 호환됩니다.[37] NanoSAM은 원래 SAM에 비해 효율성이 크게 향상되어 더 넓은 범위의 응용 프로그램에 실용적입니다.
- 실제 응용 분야 및 사용 사례: 모바일 앱 및 에지 장치에서 실시간 객체 감지 및 분할에 적합합니다.[37] 로봇 공학에서 인식 작업에 잠재적으로 사용됩니다.[14] 제로샷 개방형 어휘 인스턴스 분할을 위해 OWL-ViT와 같은 객체 감지기와 통합됩니다.[14] NanoSAM은 계산 리소스가 제한적이고 실시간 처리가 필요한 응용 프로그램에서 고급 분할 기능을 활성화합니다.
- 성능 지표 및 평가: NVIDIA Jetson Xavier NX 및 T4 GPU에서 대기 시간 및 처리량 측정값을 제공합니다.[35] COCO 2017 유효성 검사 데이터 세트에서 달성한 정확도(mIoU)를 언급합니다.[17] 속도와 크기를 원래 SAM 및 MobileSAM과 비교합니다.[17] 특정 에지 장치의 성능 지표는 개발자에게 귀중한 정보를 제공합니다.

2.11. grounding dino

- 개념적 개요 및 목표: Grounding DINO는 Transformer 기반 감지기 DINO와 접지된 사전 학습을 결합한 개방형 객체 감지기로 소개되었습니다.[38] 주요 목표는 범주 이름 또는 참조 표현과 같은 사람의 입력을 사용하여 임의의 객체를 감지하는 것입니다.[38] 제로샷 객체 감지 기능이 강조됩니다.[38] Grounding DINO는 언어를 활용하여 훨씬 더 넓은 범위의 객체를 감지할 수 있도록 합니다.
- 핵심 기능 및 작동 원리: 아키텍처에는 이미지 백본, 텍스트 백본, 이미지-텍스트

융합을 위한 기능 향상기, 언어 안내 쿼리 선택 모듈 및 교차 모달리티 디코더가 포함됩니다.[39] 이미지 및 텍스트 특징을 추출하고 융합하며 이미지 특징에서 쿼리를 선택하고 이러한 쿼리를 사용하여 객체 상자와 해당 구문을 예측하는 방법을 설명합니다.[39] 많은 양의 이미지-텍스트 데이터로 학습됩니다.[44] 언어와 시각 양식의 긴밀한 융합이 핵심입니다.

- 고유한 특징 및 변형: 개방형 객체 감지 기능이 강조됩니다.[38] COCO 및 ODinW와 같은 벤치마크에서 강력한 제로샷 성능을 강조합니다.[38] 향상된 성능과 기능을 갖춘 Grounding DINO 1.5 및 Grounding DINO 1.5 Pro와 같은 다양한 버전이 있습니다.[39] 리소스가 제한된 장치를 위한 에지 버전(EfficientViT-L1 백본)을 언급합니다.[39] 다양한 버전은 성능, 정확성 및 컴퓨팅 리소스 측면에서 다양한 요구 사항을 충족합니다.
- 실제 응용 분야 및 사용 사례: 자율 주행(알려진 객체와 알 수 없는 객체 감지).[39] 감시 및 보안(텍스트 프롬프트에 따라 특정 항목 감지).[39] 로봇 공학(새로운 객체를 인식하여 동적 환경에서 작동).[39] 이미지 검색 및 객체 추적.[41] Grounded-SAM과 같은 작업을 위해 SAM과 같은 분할 모델과 통합.[28] Grounding DINO의 언어 기반 새로운 객체 감지 기능은 다양한 응용 분야를 열어줍니다.
- 성능 지표 및 평가: COCO 제로샷에서 52.5 AP를 달성한 것을 언급합니다.[38] ODinW 제로샷 벤치마크에서 최고 기록을 세웠습니다.[38] ViT-L 백본과 더 큰 데이터 세트로 Grounding DINO 1.5 Pro의 향상된 성능을 강조합니다.[39] 어려운 벤치마크에서 강력한 성능 지표는 개방형 객체 감지에 대한 Grounding DINO의 효과를 입증합니다.

2.12. grounded-sam-2

- 개념적 개요 및 목표: Grounded SAM 2는 Grounding DINO, Florence-2 및 SAM 2를 결합하여 비디오에서 모든 것을 접지하고 추적하기 위한 기본 모델 파이프라인으로 소개되었습니다.[47] 주요 목표는 개방형 감지 및 분할 기능을 비디오 도메인으로 확장하여 언어 프롬프트에 따라 모든 객체를 추적할 수 있도록 하는 것입니다.[48] SAM 2의 비디오 분할 기능을 통합하여 원래 Grounded SAM을 기반으로 합니다.[48] Grounded SAM 2는 강력한 비디오 이해 및 상호 작용을 위한 중요한 발전입니다.
- 핵심 기능 및 작동 원리: 비디오 프레임에서 개방형 어휘 객체 감지를 위해 Grounding DINO(또는 Grounding DINO 1.5, DINO-X와 같은 변형)를 사용합니다.[48] 비디오 프레임 전체에서 감지된 객체의 고정밀 분할을 위해 SAM 2를 활용합니다.[47] 밀집 영역 캡션 및 향상된 접지를 위해 Florence-2를 사용할 가능성을 언급합니다.[48] 텍스트 프롬프트로 모든 것을 접지하고 분할하며 비디오 전체에서 객체를 추적할 수 있습니다.[48] 감지를 위한 Grounding DINO와 분할을 위한 SAM 2의 순차적 적용을 통해 모델은 언어를 기반으로 관심 객체를 식별한 다음 정확하게 경계를 구분하고 시간 경과에 따라 추적할 수 있습니다.
- 고유한 특징 및 변형: 개방형 어휘 비디오 객체 분할 및 추적 기능을 강조합니다.[48] Grounding DINO, Grounding DINO 1.5, Florence-2 및 DINO-X와 같은 다양한 접지 모델을 지원합니다.[48] SAM 2.1과의 호환성 및 고해상도 이미지를 위한 SAHI(Slicing

Aided Hyper Inference) 지원을 언급합니다.[48] 접지 모델 선택의 유연성은 사용자에게 가장 적합한 모델을 선택할 수 있도록 합니다.

- 실제 응용 분야 및 사용 사례: 언어 기반 객체 추적을 통한 비디오 분석, 감시 및 보안. 동적 환경에서 언어 안내 조작 및 상호 작용을 위한 로봇 공학.[46] 자연어 설명을 기반으로 한 자동 비디오 주석 및 편집. Grounded SAM 2는 로봇의 인식 능력을 크게 향상시킬 수 있습니다.
- 성능 지표 및 평가: 스니펫은 아키텍처 및 기능에 주로 중점을 둡니다. 성능 지표는 관련 연구 논문(arxiv.org/abs/2401.14159)에서 확인할 수 있습니다. 성능 평가는 다른 비디오 이해 모델과 비교하는 데 매우 중요합니다.

2.13. openvla

- 개념적 개요 및 목표: OpenVLA는 로봇 조작을 위한 오픈 소스 비전-언어-액션(VLA) 모델로 소개되었습니다.[50] 주요 목표는 시각적 입력과 자연어 지침에 따라 로봇이 광범위한 조작 작업을 수행할 수 있도록 하는 것입니다.[50] 사전 학습된 Prismatic-7B VLM을 대규모 로봇 조작 궤적 데이터 세트로 미세 조정하여 학습됩니다.[50] OpenVLA는 VLA 기능을 연구 커뮤니티에서 사용할 수 있도록 합니다.
- 핵심 기능 및 작동 원리: 융합된 시각적 인코더(SigLIP 및 DinoV2 백본), 프로젝터 및 Llama 2 언어 모델을 포함하는 아키텍처를 설명합니다.[50] 이미지 입력을 패치 임베딩에 매핑하고, 언어 모델의 입력 공간으로 투영하고, 결과 토큰을 연속적인 로봇 동작으로 디코딩하는 방법을 설명합니다.[50] 광범위한 작업, 장면 및 로봇 구현을 포괄하는 Open X-Embodiment 데이터 세트에서 학습됩니다.[50] 여러 시각적 인코더(SigLIP 및 DinoV2)의 융합을 통해 모델은 시각적 장면의 다양한 측면을 캡처할 수 있습니다.
- 고유한 특징 및 변형: 지원되는 로봇 구성을 위한 제로샷 제어 기능이 강조됩니다.[52] 새로운 도메인에 대한 매개 변수 효율적인 미세 조정 지원을 강조합니다.[50] 더 적은 매개 변수로 RT-2-X와 같은 폐쇄형 모델보다 뛰어난 일반 조작 성능을 제공합니다.[50] 추론 속도와 작업 성능을 크게 향상시키는 OFT(최적화된 미세 조정) 레시피(OpenVLA-OFT)의 가용성을 언급합니다.[53] OFT의 가용성은 OpenVLA의 실용성을 더욱 향상시킵니다.
- 실제 응용 분야 및 사용 사례: 자연어 지침을 통한 직접적인 로봇 제어.[52] 즉시 사용 가능한 다중 로봇 지원.[50] 실시간 시각적 장면 이해 및 동작 생성.[52] 픽앤플레이스 작업, 어수선한 장면에서 객체 조작.[50] OpenVLA는 다양한 조작 작업을 위해 로봇과의 보다 자연스럽고 직관적인 상호 작용을 가능하게 할 잠재력을 가지고 있습니다.
- 성능 지표 및 평가: 29개 작업에서 절대 작업 성공률이 16.5% 향상되어 RT-2-X(55B)보다 성능이 뛰어납니다.[51] OpenVLA-OFT가 LIBERO 시뮬레이션 벤치마크에서 상당한 동작 생성 속도 향상으로 최고 성능을 달성한 것을 언급합니다.[53] 다중 객체 및 강력한 언어 접지 기능을 갖춘 다중 작업 환경에서 강력한 일반화 결과를 강조합니다.[51] 정량적 결과는 OpenVLA의 효율성과 효과를

입증합니다.

2.14. MobileSAMv2

- 개념적 개요 및 목표: MobileSAMv2는 원래 SAM에 비해 더 빠른 "Segment Everything"(SegEvery)을 목표로 하는 프로젝트로 소개되었습니다.[55] 주요 목표는 경쟁력 있는 성능을 유지하면서 이미지의 모든 객체를 분할하는 프로세스를 가속화하는 것입니다.[55] "Segment Anything"(SegAny)의 속도를 높이는 데 중점을 두는 MobileSAM과 차별화됩니다.[55] MobileSAMv2는 이 포괄적인 장면 이해를 리소스가 제한된 환경에서 더 실현 가능하게 만듭니다.
- 핵심 기능 및 작동 원리: 객체 인식 프롬프트 샘플링 및 프롬프트 안내 마스크 디코딩의 2단계 프레임워크를 설명합니다.[55] 객체 인식 프롬프트를 샘플링하기 위해 최신 객체 인식 네트워크(SA-1B의 하위 집합으로 학습된 YOLOv8) 사용을 설명합니다.[55] 점 프롬프트보다 효율적인 SAM 마스크 디코더에 이러한 경계 상자를 직접 프롬프트로 사용하는 것을 언급합니다.[55] 프롬프트로 사용하기 전에 겹치는 경계 상자를 필터링하기 위해 NMS(Non-Maximum Suppression) 사용을 언급합니다.[56] 핵심 혁신은 SAM을 위한 프롬프트 샘플링을 안내하기 위해 객체 감지기를 사용하는 것입니다.
- 고유한 특징 및 변형: SegEvery 작업 속도 향상에 중점을 둡니다.[55] 기존 그리드 검색 대신 객체 인식 프롬프트 샘플링을 강조합니다.[55] 효율적인 SegAny 및 SegEvery를 위한 통합 프레임워크에 기여하는 MobileSAM의 증류된 이미지 인코더와의 호환성을 언급합니다.[28] MobileSAMv2는 원래 SAM에 비해 더 지능적이고 효율적인 방법을 제공합니다.
- 실제 응용 분야 및 사용 사례: 자율 시스템의 장면 이해, 포괄적인 이미지 분석 및 고급 이미지 편집과 같이 이미지의 모든 객체 분할이 필요한 응용 프로그램. 로봇 폐기물 분류에 잠재적으로 사용됩니다.[57] MobileSAMv2는 다양한 도메인에서 시각적 장면의 보다 상세하고 자동화된 분석을 가능하게 할 수 있습니다.
- 성능 지표 및 평가: LVIS 데이터 세트에서 제로샷 객체 제안에 대해 평균 성능이 3.6% 향상된 것을 언급합니다.[28] 원래 SAM에 비해 더 빠른 SegEvery를 달성하면서 경쟁력 있는 성능을 제공합니다.[55] 성능 지표는 MobileSAMv2가 정확도를 크게 희생하지 않고 더 빠른 SegEvery라는 목표를 달성했음을 나타냅니다.

2.15. ESAM

- 개념적 개요 및 목표: 스니펫은 ESAM에 대한 적어도 세 가지 가능한 의미를 나타냅니다.
  - 임의 마스크의 효율적인 합산: 천체 물리학에서 탈분산에 사용되는 많은 임의 2D 마스크의 효율적인 1D 컨볼루션을 위한 새로운 접근 방식입니다.[58] 이것은 컴퓨터 비전 모델의 맥락에서 의도된 의미가 아닐 가능성이 높습니다.
  - OpenEdge의 ESAM 디렉토리 구조: OpenEdge 설치 아티팩트 및 OpenEdge 루트 설치 경로의 무결성을 보호하기 위한 고정된 절대 파일 시스템 공간을

나타냅니다.[59] 이것도 의도된 의미가 아닐 가능성이 높습니다.
    - 디지털 카메라의 자동 노출 알고리즘: 디지털 카메라에서 셔터 속도를 계산하는 데 사용되는 알고리즘입니다.[60] 이는 이미지 처리와 더 관련이 있지만 목록에 있는 다른 컴퓨터 비전 모델과 같은 일반적인 모델은 아닙니다. 다른 나열된 항목의 맥락을 고려할 때 디지털 카메라의 자동 노출과 관련된 의미가 가장 plausibly 보입니다.
    - 핵심 기능 및 작동 원리: ESAM이 자동 노출 알고리즘을 의미하는 경우 핵심 기능은 장면의 조명 조건에 따라 이미지를 캡처하기 위한 적절한 셔터 속도를 계산하는 것입니다.[60] 적절한 노출 설정을 결정하기 위해 원시 이미지에서 검은색 및 흰색 픽셀의 백분율을 분석할 가능성이 높습니다.[60]
    - 고유한 특징 및 변형: 분할 또는 객체 감지와 같은 획득 후 분석보다는 이미지 획득에 중점을 둔 알고리즘입니다.
    - 실제 응용 분야 및 사용 사례: 최적의 이미지 밝기를 위해 디지털 카메라에서 노출 매개 변수를 자동으로 설정하는 데 사용됩니다.[60]
    - 성능 지표 및 평가: 성능은 적절한 노출 및 디테일 보존 측면에서 결과 이미지의 품질을 기준으로 평가될 가능성이 높습니다.

3. 비교 분석 및 상호 연결

- 관계 매핑:
    - 기초 모델 및 확장: SAM2는 SAM을 기반으로 비디오로 기능을 확장합니다. Grounding DINO는 DINO 객체 감지기를 언어 이해로 확장합니다. CLIP-SEG는 분할을 위해 CLIP을 기반으로 합니다. CLIP-RT는 로봇 동작 예측을 위해 CLIP을 확장합니다.
    - 효율성 중심: NanoOWL과 NanoSAM은 각각 에지 배포 및 실시간 성능을 위해 OWL-ViT와 SAM의 최적화된 버전입니다. MobileSAMv2는 모든 것을 분할하기 위해 SAM의 효율성을 개선하는 데 중점을 둡니다.
    - 의미 통합: Semantic-SAM과 Semantic-Segment-Anything의 개념은 SAM과 같은 분할 모델에 의미 이해를 추가하는 것을 목표로 합니다. ZegCLIP도 분할을 위해 CLIP의 의미 이해를 활용합니다. Grounded SAM 2는 Grounding DINO의 감지 기능과 SAM 2의 분할을 결합하여 비디오에서 의미 정보 및 추적을 제공합니다.
    - 비전-언어-액션 모델: CLIP-RT와 OpenVLA는 로봇 제어를 위해 비전, 언어 및 동작을 통합하는 별도의 모델 범주를 나타냅니다.
    - 이상치: EVLA는 목록에 있는 컴퓨터 비전 모델의 핵심 주제와 관련이 없어 보이며, 다른 도메인의 기술 또는 이 맥락에서 덜 일반적인 용어를 나타낼 수 있습니다. ESAM은 이미지 캡처와 관련이 있지만 다른 모델의 분할 및 객체 감지 작업과는 직접적인 관련이 없습니다.
- 기능 비교표:

특징	SAM2	ZegCLIP	SCLIP	CLIP-SEG	NanoOWL	Semantic-SAM	Semantic-Segment-Anything	EVLA	CLIP-RT	NanoSAM	GroundingDINO	Grounded-SAM-2	OpenVLA	MobileSAMv2	ESAM(자동노출)
작업	이미지/비디오분할	제로샷의미분할	준지도비전-언어사전학습	제로/원샷이미지분할	개방형어휘객체감지	세분화된의미분할	의미및인스턴스분할	의료치료/전파간섭계	언어조건부로봇동작	가속화된이미지분할	개방형객체감지	개방형어휘비디오분할및추적	언어조건부로봇동작	더빠른모든것분할	자동노출제어
입력	이미지,비디오,프롬프트	이미지,텍스트	이미지-텍스트쌍,쌍을이	이미지,텍스트/이미지프	이미지,텍스트	이미지,점/상자프롬프	이미지	레이저에너지/전파	이미지,텍스트지침	이미지,점/상자프롬프	이미지,텍스트	이미지,비디오,텍스트	이미지,언어지침	이미지,객체감지기출력	이미지

	(클릭, 상자, 마스크)		루지않은이미지	롬프트		트				트				
아키텍처	메모리포함트랜스포머	경량디코더포함CLIP기반	의사레이블링포함CLIP기반	CLIP+트랜스포머디코더	TensorRT포함최적화된OWL-ViT	텍스트인코더포함MaskDINO기반	다양함(종종SAM기반)	다양함(레이저시스템/전파망원경배열)	CLIP기반트랜스포머	TensorRT포함증류된SAM	트랜스포머기반(DINO+GLIP)	Grounding DINO+SAM2+Florence-2	VLM(DINOv2, SigLIP, Llama 2)	YOLOv8+SAM디코더	픽셀값분석알고리즘
제로샷가능	예	예	아니요	예	예	예	예	해당없음	예	예	예	예	예	아니요	예
실시간성능	예(비디오	가능성있음	-	가능성있음	예(Jetson	가능성있음	-	-	가능성있음	예(Jetson	-	-	예(OFT포	가능성있음	예

	약 44 FPS)				Orin에서)					Orin에서)			함)	(속도최적화)	
에지배포	예	가능성있음	-	가능성있음	예(Jetson에최적화됨)	가능성있음	-	-	-	예(Jetson에최적화됨)	가능성있음(에지버전사용가능)	-	-	예(MobileSAM파트)	예(카메라에서)

4. 현재 상황 및 미래 방향

- 분야에서의 위치: SAM 및 그 변형(SAM2, NanoSAM, MobileSAMv2)은 프롬프트 가능성 및 제로샷 기능으로 인해 연구 및 응용 분야에서 널리 채택되어 범용 이미지 및 비디오 분할을 위한 기본 모델이 되었습니다. CLIP 및 그 변형(ZegCLIP, SCLIP, CLIP-SEG, CLIP-RT, Grounding DINO, Grounded SAM 2, OpenVLA)과 같은 비전-언어 모델은 연구의 최전선에 있으며, 보다 유연하고 의미적으로 풍부한 시각적 이해 및 상호 작용을 가능하게 합니다. 이는 시각적 작업에 대한 언어 지침을 이해하고 응답할 수 있는 모델로의 추세를 나타냅니다. 에지 장치에 최적화된 모델(NanoOWL, NanoSAM, MobileSAMv2)은 컴퓨팅 리소스가 제한된 실제 시나리오에서 고급 컴퓨터 비전을 배포하는 데 매우 중요하며, 장치 내 AI에 대한 수요 증가를 반영합니다. "Semantic-Segment-Anything"의 개념은 클래스에 구애받지 않는 분할과 의미 이해 사이의 간극을 메우는 것을 목표로 하는 분야의 핵심 방향을 나타냅니다. 현재 상황은 광범위한 기능을 갖춘 기본 모델, 시각 작업에 언어 이해 통합, 에지 장치에 효율적인 배포에 대한 관심 증가라는 특징이 있습니다. 많은 모델 이름에 "SAM"과 "CLIP"이 있다는 것은 그들의 기초적인 중요성을 나타냅니다. "Nano" 및 "Mobile" 버전의 출현은 효율성에 대한 초점을 강조합니다. 여러 모델에서 "Semantic"이라는 용어가 명시적으로 언급된 것은 의미 이해의

중요성을 나타냅니다.

- 새로운 트렌드 및 잠재적 발전: 보다 효율적이고 정확한 분할 및 객체 감지를 위한 기본 모델의 추가 개발. 보다 정교한 장면 이해, 추론 및 상호 작용을 위한 시각 및 언어 통합 증가. 대규모 모델을 에지 장치에 배포하기 위한 증류 및 TensorRT와 같은 최적화 기술에 대한 지속적인 집중. 향상된 3D 이해를 위해 깊이 정보와 같은 RGB 이미지 외에 더 넓은 범위의 양식을 처리할 수 있는 모델 개발.[47] 실제 환경에서 복잡한 지침을 이해하고 실행할 수 있는 보다 유능하고 다재다능한 로봇을 위한 비전-언어-액션 모델의 발전. 다양한 작업(감지, 분할, 캡션 생성)의 통합 모델로의 잠재적 수렴. 이러한 영역에서 컴퓨터 비전의 미래는 다양한 환경에서 효과적으로 작동할 수 있는 보다 강력하고 효율적이며 다재다능한 모델을 포함할 가능성이 높습니다. 현재 모델에서 관찰된 추세(기본 모델, 언어 통합, 에지 최적화, 다중 양식)는 미래에 계속되고 가속화될 가능성이 높습니다.

5. 결론

이 보고서는 SAM2, ZegCLIP, SCLIP, CLIP-SEG, NanoOWL, Semantic-SAM, Semantic-Segment-Anything, EVLA, CLIP-RT, NanoSAM, grounding dino, grounded-sam-2, openvla, MobileSAMv2 및 ESAM을 포함한 15가지 최신 컴퓨터 비전 모델 및 기술에 대한 심층 분석을 제공했습니다. 각 기술의 개념, 핵심 기능, 고유한 특징, 실제 응용 분야 및 성능 지표를 자세히 살펴보았습니다. 비교 분석을 통해 이러한 기술 간의 관계와 차이점을 강조하고, 기본 모델, 효율성 중심 접근 방식, 의미 통합 및 비전-언어-액션 모델의 주요 추세를 보여주었습니다. 또한 에지 배포를 위해 최적화된 모델의 중요성이 점점 커지고 있음을 강조했습니다. 마지막으로, 이 보고서는 이러한 발전이 컴퓨터 비전 분야의 미래와 다양한 응용 분야에 갖는 중요성에 대한 관점을 제공합니다.

6. 참고 문헌

.[1]

참고 자료

1. Deep-diving into SAM 2: How Quality Data Propelled Their Visual Segmentation Model, 4월 25, 2025에 액세스, https://kili-technology.com/data-labeling/deep-diving-into-sam2-how-quality-data-propelled-meta-s-visual-segmentation-model
2. SAM 2: Meta's Next-Gen Model for Video and Image Segmentation | DigitalOcean, 4월 25, 2025에 액세스, https://www.digitalocean.com/community/tutorials/sam-2-metas-next-gen-model-for-video-and-image-segmentation
3. Meta's SAM-2: The Future of Real-Time Visual Segmentation - Analytics Vidhya, 4월 25, 2025에 액세스,

https://www.analyticsvidhya.com/blog/2024/08/meta-sam-2/
4. Introducing Meta Segment Anything Model 2 (SAM 2), 4월 25, 2025에 액세스, https://ai.meta.com/sam2/
5. Semantic intelligence meets photogrammetry in PIX4Dmatic | Pix4D, 4월 25, 2025에 액세스, https://www.pix4d.com/labs/semantic-intelligence-photogrammetry-pix4dmatic
6. sam2 Model by Meta - NVIDIA NIM APIs, 4월 25, 2025에 액세스, https://build.nvidia.com/meta/sam2/modelcard
7. Relationship Prompt Learning is Enough for Open-Vocabulary Semantic Segmentation - NIPS papers, 4월 25, 2025에 액세스, https://papers.nips.cc/paper_files/paper/2024/file/8773cdaf02c5af3528e05f1cee816129-Paper-Conference.pdf
8. ZegCLIP: Towards Adapting CLIP for Zero-shot Semantic Segmentation - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/373316526_ZegCLIP_Towards_Adapting_CLIP_for_Zero-shot_Semantic_Segmentation
9. S-CLIP: Semi-supervised Vision-Language Pre-training using Few Specialist Captions - GitHub, 4월 25, 2025에 액세스, https://github.com/alinlab/s-clip
10. CLIPSeg - Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/docs/transformers/main/model_doc/clipseg
11. Fine-Tuning Clipseg Techniques | Restackio, 4월 25, 2025에 액세스, https://www.restack.io/p/fine-tuning-answer-clipseg-techniques-cat-ai
12. Zero-shot Image Segmentation with CLIPSeg - Orchestra, 4월 25, 2025에 액세스, https://www.getorchestra.io/guides/zero-shot-image-segmentation-with-clipseg
13. Applications of ClipSeg Model | Restackio, 4월 25, 2025에 액세스, https://www.restack.io/p/context-aware-ai-application-techniques-answer-clipseg-applications-cat-ai
14. NVIDIA-AI-IOT/nanoowl: A project that optimizes OWL-ViT for real-time inference with NVIDIA TensorRT. - GitHub, 4월 25, 2025에 액세스, https://github.com/NVIDIA-AI-IOT/nanoowl
15. NanoOWL - NVIDIA Jetson AI Lab, 4월 25, 2025에 액세스, https://www.jetson-ai-lab.com/vit/tutorial_nanoowl.html
16. 157T Flagship AI Box - PowerPoint 演示文稿, 4월 25, 2025에 액세스, https://download.t-firefly.com/%E4%BA%A7%E5%93%81%E8%A7%84%E6%A0%BC%E6%96%87%E6%A1%A3/%E5%B5%8C%E5%85%A5%E5%BC%8F%E4%B8%BB%E6%9C%BA/AIBOX-OrinNano%26AIBOX-OrinNX_157T%20Flagship%20AI%20Box_Product%20Introduction.pdf?v=1741929302
17. NVIDIA-AI-IOT/nanosam: A distilled Segment Anything (SAM) model capable of running real-time with NVIDIA TensorRT - GitHub, 4월 25, 2025에 액세스, https://github.com/NVIDIA-AI-IOT/nanosam
18. NVIDIA Jetson Orin Nano Developer Kit - Seeed Studio, 4월 25, 2025에 액세스, https://files.seeedstudio.com/wiki/Jetson-Orin-Nano-DevKit/jetson-orin-nano-developer-kit-datasheet.pdf
19. Make Your Existing NVIDIA Jetson Orin Devices Faster with Super Mode, 4월 25, 2025에 액세스,

https://www.edge-ai-vision.com/2025/02/make-your-existing-nvidia-jetson-orin-devices-faster-with-super-mode/
20. Make your existing NVIDIA® Jetson Orin™ devices faster with Super Mode - e-con Systems, 4월 25, 2025에 액세스, https://www.e-consystems.com/blog/camera/products/make-your-existing-nvidia-jetson-orin-devices-faster-with-super-mode/
21. Segment and Recognize Anything at Any Granularity - European Computer Vision Association, 4월 25, 2025에 액세스, https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/06495.pdf
22. Semantic-SAM: Segment and Recognize Anything at Any Granularity - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2307.04767
23. eastoc/ComfyUI_SemanticSAM - GitHub, 4월 25, 2025에 액세스, https://github.com/eastoc/ComfyUI_SemanticSAM
24. Semantic segmentation: Complete guide [Updated 2024] | SuperAnnotate, 4월 25, 2025에 액세스, https://www.superannotate.com/blog/guide-to-semantic-segmentation
25. Segment Anything Model (SAM) - The Complete 2025 Guide - viso.ai, 4월 25, 2025에 액세스, https://viso.ai/deep-learning/segment-anything-model-sam-explained/
26. A Hands-on Guide to Using Segment Anything - Labelvisor, 4월 25, 2025에 액세스, https://www.labelvisor.com/effortless-segmentation-in-seconds-a-hands-on-guide-to-using-segment-anything/
27. segment anything detailed guide | ComfyUI - RunComfy, 4월 25, 2025에 액세스, https://www.runcomfy.com/comfyui-nodes/comfyui_segment_anything
28. MobileSAMv2: Faster Segment Anything to Everything - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2312.09579
29. Endovenous laser ablation (EVLA): a review of mechanisms, modeling outcomes, and issues for debate - PMC, 4월 25, 2025에 액세스, https://pmc.ncbi.nlm.nih.gov/articles/PMC3953603/
30. MULTI-FREQUENCY-SYNTHESIS AND WIDE-FIELD IMAGING WITH THE EVLA - URSI, 4월 25, 2025에 액세스, https://www.ursi.org/proceedings/procGA11/ursi/J06-5.pdf
31. EVLA Memo 67 W PROJECTION: A NEW ALGORITHM FOR NON-COPLANAR BASELINES - NRAO Library, 4월 25, 2025에 액세스, https://library.nrao.edu/public/memos/evla/EVLAM_67.pdf
32. gicheonkang/clip-rt: + CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision - GitHub, 4월 25, 2025에 액세스, https://github.com/gicheonkang/clip-rt
33. (PDF) CLIP-RT: Learning Language-Conditioned Robotic Policies from Natural Language Supervision - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/385510350_CLIP-RT_Learning_Language-Conditioned_Robotic_Policies_from_Natural_Language_Supervision
34. NanoSAM - NVIDIA Jetson AI Lab, 4월 25, 2025에 액세스, https://www.jetson-ai-lab.com/vit/tutorial_nanosam.html

35. dragonSwing/nanosam - Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/dragonSwing/nanosam
36. zz990099/EasyDeploy: This project includes implementations of YOLOv8, RT-DETR-V2(RTDETR), MobileSAM, and NanoSAM on TensorRT, ONNX Runtime, and RKNN, along with support for asynchronous inference workflows. It provides a user-friendly deep learning deployment tool for seamless algorithm migration across different inference frameworks. - GitHub, 4월 25, 2025에 액세스, https://github.com/zz990099/EasyDeploy
37. MobileSAM (Mobile Segment Anything Model) - Ultralytics YOLO Docs, 4월 25, 2025에 액세스, https://docs.ultralytics.com/models/mobile-sam/
38. Grounding DINO - Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/docs/transformers/en/model_doc/grounding-dino
39. Grounding DINO 1.5: Pushing the Boundaries of Open-Set Object Detection | DigitalOcean, 4월 25, 2025에 액세스, https://www.digitalocean.com/community/tutorials/grounding-dino-1-5-open-set-object-detection
40. Grounding DINO - Breaking Boundaries in Object Detection - Eizen AI, 4월 25, 2025에 액세스, https://eizen.ai/groundingDino.html
41. Zero-shot Object Detection Using Grounding DINO Base - Analytics Vidhya, 4월 25, 2025에 액세스, https://www.analyticsvidhya.com/blog/2024/10/grounding-dino-base/
42. Grounded-SAM Explained: A New Image Segmentation Paradigm? - viso.ai, 4월 25, 2025에 액세스, https://viso.ai/deep-learning/grounded-sam/
43. IDEA-Research/GroundingDINO: [ECCV 2024] Official implementation of the paper "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection" - GitHub, 4월 25, 2025에 액세스, https://github.com/IDEA-Research/GroundingDINO
44. Grounding DINO - NGC Catalog - NVIDIA, 4월 25, 2025에 액세스, https://catalog.ngc.nvidia.com/orgs/nvidia/teams/tao/models/grounding_dino
45. Grounding DINO : SOTA Zero-Shot Object Detection - Roboflow Blog, 4월 25, 2025에 액세스, https://blog.roboflow.com/grounding-dino-zero-shot-object-detection/
46. IDEA-Research/Grounded-Segment-Anything: Grounded SAM: Marrying Grounding DINO with Segment Anything & Stable Diffusion & Recognize Anything - Automatically Detect , Segment and Generate Anything - GitHub, 4월 25, 2025에 액세스, https://github.com/IDEA-Research/Grounded-Segment-Anything
47. Grounded SAM 2 - GRID Documentation, 4월 25, 2025에 액세스, https://docs.scaledfoundations.ai/models/segmentation/gsam2
48. IDEA-Research/Grounded-SAM-2: Grounded SAM 2: Ground and Track Anything in Videos with Grounding DINO, Florence-2 and SAM 2 - GitHub, 4월 25, 2025에 액세스, https://github.com/IDEA-Research/Grounded-SAM-2
49. [P] Grounded SAM 2: Ground and Track Anything : r/MachineLearning - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/MachineLearning/comments/1elmxnq/p_grounded_sam_2_ground_and_track_anything/

50. OpenVLA: An Open-Source Vision-Language-Action Model, 4월 25, 2025에 액세스, https://openvla.github.io/
51. OpenVLA: An Open-Source Vision-Language-Action Model - GitHub, 4월 25, 2025에 액세스, https://raw.githubusercontent.com/mlresearch/v270/main/assets/kim25c/kim25c.pdf
52. openvla-7b - PromptLayer, 4월 25, 2025에 액세스, https://www.promptlayer.com/models/openvla-7b
53. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2502.19645v1
54. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success, 4월 25, 2025에 액세스, https://openvla-oft.github.io/
55. MobileSAMv2: Faster Segment Anything to Everything - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2312.09579v1
56. (PDF) MobileSAMv2: Faster Segment Anything to Everything - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/376579294_MobileSAMv2_Faster_Segment_Anything_to_Everything
57. Versatile waste sorting in small batch and flexible manufacturing industries using deep learning techniques, 4월 25, 2025에 액세스, https://pmc.ncbi.nlm.nih.gov/articles/PMC11782652/
58. [2412.10678] Efficient Summation of Arbitrary Masks -- ESAM - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/abs/2412.10678
59. ESAM directory structure - Progress Documentation, 4월 25, 2025에 액세스, https://docs.progress.com/bundle/openedge-security-and-auditing/page/ESAM-directory-structure.html
60. On the left: shutter speed calculated by ESAM algorithm, percentage of... - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/figure/On-the-left-shutter-speed-calculated-by-ESAM-algorithm-percentage-of-black-white_fig9_338751295
61. [Literature Review] RGBD Objects in the Wild: Scaling Real-World 3D Object Learning from RGB-D Videos - Moonlight, 4월 25, 2025에 액세스, https://www.themoonlight.io/en/review/rgbd-objects-in-the-wild-scaling-real-world-3d-object-learning-from-rgb-d-videos
62. RGB-D Cube R-CNN: 3D Object Detection with Selective Modality Dropout - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/384417771_RGB-D_Cube_R-CNN_3D_Object_Detection_with_Selective_Modality_Dropout
63. A RGB-D Semantic Dense SLAM Based on 3D Multi Level Pyramid Gaussian Splatting, 4월 25, 2025에 액세스, https://arxiv.org/html/2412.01217v1
64. Articulated Object Manipulation using Online Axis Estimation with SAM2-Based Tracking, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.16287v1
65. RGBDS-SLAM: A RGB-D Semantic Dense SLAM Based on 3D Multi Level Pyramid Gaussian Splatting - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/386375524_RGBDS-SLAM_A_RGB-D_S

emantic_Dense_SLAM_Based_on_3D_Multi_Level_Pyramid_Gaussian_Splatting
66. A Pipeline for Segmenting and Structuring RGB-D Data for Robotics Applications - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2410.17988v1
67. Examples - Rerun, 4월 25, 2025에 액세스, https://rerun.io/examples
68. IntelligentRoboticsLabs/yolact_ros_3d - GitHub, 4월 25, 2025에 액세스, https://github.com/IntelligentRoboticsLabs/yolact_ros_3d
69. Pick and Place Using MoveIt 2 and Perception – ROS 2 Jazzy - Automatic Addison, 4월 25, 2025에 액세스, https://automaticaddison.com/pick-and-place-task-using-moveit-2-and-perception-ros2-jazzy/
70. IntelligentRoboticsLabs/gb_visual_detection_3d - GitHub, 4월 25, 2025에 액세스, https://github.com/IntelligentRoboticsLabs/gb_visual_detection_3d
71. OpenNav: Efficient Open Vocabulary 3D Object Detection for Smart Wheelchair Navigation, 4월 25, 2025에 액세스, https://arxiv.org/html/2408.13936v1
72. yolo_ros/README.md at main - GitHub, 4월 25, 2025에 액세스, https://github.com/mgonzs13/yolov8_ros/blob/main/README.md
73. ethz-asl/rgbd_segmentation: ROS package for geometric-semantic segmentation of RGB-D sequences. - GitHub, 4월 25, 2025에 액세스, https://github.com/ethz-asl/rgbd_segmentation
74. How should I add semantic segmentation data to my costmap? - ROS Answers archive, 4월 25, 2025에 액세스, http://answers.ros.org/question/332477/
75. Grounded Object Segmentation and Localization with Gaussian Splatting SLAM - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.16944v1
76. Object Detection with ROS2 | ROS2 Developers Open Class #140 - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=CajBysjKKUk
77. Aruco Pose Estimation with ROS2, using RGB and Depth camera images from Realsense D435 - GitHub, 4월 25, 2025에 액세스, https://github.com/AIRLab-POLIMI/ros2-aruco-pose-estimation
78. Generate Synthetic Data for Deep Object Pose Estimation Training with NVIDIA Isaac ROS, 4월 25, 2025에 액세스, https://developer.nvidia.com/blog/generate-synthetic-data-for-deep-object-pose-estimation-training-with-nvidia-isaac-ros/
79. Pose Estimation — isaac_ros_docs documentation, 4월 25, 2025에 액세스, https://nvidia-isaac-ros.github.io/concepts/pose_estimation/index.html
80. Isaac ROS Pose Estimation — isaac_ros_docs documentation, 4월 25, 2025에 액세스, https://nvidia-isaac-ros.github.io/repositories_and_packages/isaac_ros_pose_estimation/index.html
81. S-Clip E2: A New Concept of Clipping Algorithms - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/profile/Vaclav-Skala/publication/262216436_S-clip_E2_A_new_concept_of_clipping_algorithms/links/5486e4050cf289302e2d89f7/S-clip-E2-A-new-concept-of-clipping-algorithms.pdf
82. Semantic Segmentation Algorithm - Amazon SageMaker AI - AWS Documentation, 4월 25, 2025에 액세스,

https://docs.aws.amazon.com/sagemaker/latest/dg/semantic-segmentation.html
83. CLIP Explained - Papers With Code, 4월 25, 2025에 액세스, https://paperswithcode.com/method/clip
84. NanoScan Lab - Scienta Omicron, 4월 25, 2025에 액세스, https://scientaomicron.com/en/products-solutions/electron-spectroscopy/NanoScan-Lab
85. Grounding DINO - NVIDIA Docs, 4월 25, 2025에 액세스, https://docs.nvidia.com/tao/tao-toolkit/text/cv_finetuning/pytorch/object_detection/grounding_dino.html
86. Lite-SAM Is Actually What You Need for Segment Everything - European Computer Vision Association, 4월 25, 2025에 액세스, https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/05077.pdf
87. CLIP-driven Dual Feature Enhancing Network for Gaze Estimation - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2502.20128v1
88. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2502.14786v1
89. sCLIP—an integrated platform to study RNA–protein interactomes in biomedical research: identification of CSTF2tau in alternative processing of small nuclear RNAs - PubMed Central, 4월 25, 2025에 액세스, https://pmc.ncbi.nlm.nih.gov/articles/PMC5449641/
90. CLIP — NVIDIA NeMo Framework User Guide, 4월 25, 2025에 액세스, https://docs.nvidia.com/nemo-framework/user-guide/latest/nemotoolkit/multimodal/vlm/clip.html
91. Vision Language Action Models (VLA) Overview: LeRobot Policies Demo, 4월 25, 2025에 액세스, https://learnopencv.com/vision-language-action-models-lerobot-policy/
92. Guide to Vision-Language Models (VLMs) - Encord, 4월 25, 2025에 액세스, https://encord.com/blog/vision-language-models-guide/
93. New algorithm unlocks high-resolution insights for computer vision | MIT News, 4월 25, 2025에 액세스, https://news.mit.edu/2024/featup-algorithm-unlocks-high-resolution-insights-computer-vision-0318
94. GeForce RTX 50 series - Wikipedia, 4월 25, 2025에 액세스, https://en.wikipedia.org/wiki/GeForce_RTX_50_series
95. Gaming on 2nd Gen RTX Architecture | GeForce RTX 3080 - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=wkO5NM5G2Zo
96. meta / sam2 - NVIDIA API Documentation, 4월 25, 2025에 액세스, https://docs.api.nvidia.com/nim/reference/meta-sam2
97. Computer Vision Algorithms | GeeksforGeeks, 4월 25, 2025에 액세스, https://www.geeksforgeeks.org/computer-vision-algorithms/
98. Machine learning in image-based metal additive manufacturing process monitoring and control: A review - AccScience Publishing, 4월 25, 2025에 액세스, https://accscience.com/journal/ESAM/1/1/10.36922/esam.8548
99. System Implementation – MAGIC - MRSD Projects, 4월 25, 2025에 액세스,

https://mrsdprojects.ri.cmu.edu/2025teamh/system-implementation/
100. zdata-inc/sam2_realtime: The repository provides code for running inference with the Meta Segment Anything Model 2 (SAM 2) in real-time. - GitHub, 4월 25, 2025에 액세스, https://github.com/zdata-inc/sam2_realtime
101. NVIDIA-AI-IOT/ROS2-NanoOWL: ROS 2 node for open-vocabulary object detection using NanoOWL. - GitHub, 4월 25, 2025에 액세스, https://github.com/NVIDIA-AI-IOT/ROS2-NanoOWL
102. OWLv2 - Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/docs/transformers/model_doc/owlv2
103. jetson-containers/packages/vit/nanoowl/Dockerfile at master - GitHub, 4월 25, 2025에 액세스, https://github.com/dusty-nv/jetson-containers/blob/master/packages/vit/nanoowl/Dockerfile
104. Creating a Point Cloud Using LIO-SAM with ROS2 and Gazebo - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=NNR9RUNz5Pg
105. RMP-SAM: Towards Real-Time Multi-Purpose Segment Anything | OpenReview, 4월 25, 2025에 액세스, https://openreview.net/forum?id=1pXzC30ry5
106. SAM-Automatic-Semantic-Segmentation - Kaggle, 4월 25, 2025에 액세스, https://www.kaggle.com/code/yogendrayatnalkar/sam-automatic-semantic-segmentation
107. CLIP-RT : Learning Language-Conditioned Robotic Policies from Natural Language Supervision, 4월 25, 2025에 액세스, https://clip-rt.github.io/
108. ROS 2 on Microcontrollers with micro-ROS & Zephyr // Zephyr Tech Talk #011 - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=E8KZEuUCtVA
109. Performance Comparison of YoloX and Clip - Restack, 4월 25, 2025에 액세스, https://www.restack.io/p/lisp-performance-comparison-answer-yolox-clip-cat-ai
110. What is Clip Art? The Overview, Uses, & Benefits | Lenovo US, 4월 25, 2025에 액세스, https://www.lenovo.com/us/en/glossary/clipart/
111. ohlerlab/clip_pipeline: CLIP pipeline - GitHub, 4월 25, 2025에 액세스, https://github.com/ohlerlab/clip_pipeline
112. openai/CLIP: CLIP (Contrastive Language-Image Pretraining), Predict the most relevant text snippet given an image - GitHub, 4월 25, 2025에 액세스, https://github.com/openai/CLIP
113. Program NanoSaur robot with ROS2 | ROS Developers Live Class #130 - YouTube, 4월 25, 2025에 액세스, https://www.youtube.com/watch?v=7G4HKfIISO8
114. Exploring Text Generation on NVIDIA Jetson with Generative AI Lab - RidgeRun, 4월 25, 2025에 액세스, https://www.ridgerun.com/post/exploring-text-generation-on-nvidia-jetson-with-generative-ai-lab
115. jetson-containers - AWS, 4월 25, 2025에 액세스, https://cdck-file-uploads-global.s3.dualstack.us-west-2.amazonaws.com/nvidia/original/4X/4/f/5/4f57a904c41838f1a9a02782034c8a397270997c.txt

116. nanosam-cpp/nanosam.vcxproj.user at main · spacewalk01/nanosam-cpp - GitHub, 4월 25, 2025에 액세스, https://github.com/spacewalk01/nanosam-cpp/blob/main/nanosam.vcxproj.user
117. Rosbag2 Anonymizer - Autoware Universe Documentation, 4월 25, 2025에 액세스, https://autowarefoundation.github.io/autoware-documentation/main/datasets/data-anonymization/
118. First Look at NVIDIA Jetson Orin Nano Super - The Most Affordable Generative AI Supercomputer - DEV Community, 4월 25, 2025에 액세스, https://dev.to/ajeetraina/first-look-at-nvidia-jetson-orin-nano-super-the-most-affordable-generative-ai-supercomputer-1pe9
119. grounding-dino | AI Model Details - AIModels.fyi, 4월 25, 2025에 액세스, https://www.aimodels.fyi/models/replicate/grounding-dino-adirik
120. transformers/docs/source/en/model_doc/grounding-dino.md at main - GitHub, 4월 25, 2025에 액세스, https://github.com/huggingface/transformers/blob/main/docs/source/en/model_doc/grounding-dino.md
121. SAM2 Segmentation on Jetson AGX Orin - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/sam2-segmentation-on-jetson-agx-orin/325069
122. Grounded SAM 2 Base Model - Autodistill, 4월 25, 2025에 액세스, https://docs.autodistill.com/base_models/grounded-sam-2/
123. patrick-tssn/Streaming-Grounded-SAM-2 - GitHub, 4월 25, 2025에 액세스, https://github.com/patrick-tssn/Streaming-Grounded-SAM-2
124. Verify openvla in robosuite - Jetson AGX Orin - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/verify-openvla-in-robosuite/325858
125. OpenVLA: An Open-Source Vision-Language-Action Model, 4월 25, 2025에 액세스, https://proceedings.mlr.press/v270/kim25c.html
126. OpenVLA: An open-source vision-language-action model for robotic manipulation. - GitHub, 4월 25, 2025에 액세스, https://github.com/openvla/openvla
127. Nav 2 in ROS 2 for autonomous Navigation using SLAM for Indoor Mobile Robots - YouTube, 4월 25, 2025에 액세스, https://m.youtube.com/watch?v=GSuqO0p2mlk
128. Real-time performance tunning - Jetson AGX Orin - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/real-time-performance-tunning/329155
129. pyproject.toml - vuer-ai/feature-splatting - GitHub, 4월 25, 2025에 액세스, https://github.com/vuer-ai/feature-splatting/blob/main/pyproject.toml
130. 3D reconstruction from RGBD images. : r/computervision - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/computervision/comments/1ihj76r/3d_reconstruction_from_rgbd_images/

131. tianyang/repobench_ablation_64k · Datasets at Hugging Face, 4월 25, 2025에 액세스, https://huggingface.co/datasets/tianyang/repobench_ablation_64k/viewer
132. ZalZarak/RGBD-to-3D-Pose: This repository extracts 3D-coordinates of joint positions of a humanoid using OpenPose and a IntelRealSense Depth-Camera. With those joints it simulates a humanoid having spheres and cylinders as limbs in PyBullet. It is designed detect humans for collision avoidance for robots (proof of concept - GitHub, 4월 25, 2025에 액세스, https://github.com/ZalZarak/RGBD-to-3D-Pose
133. Language-Driven Closed-Loop Grasping with Model-Predictive Trajectory Replanning, 4월 25, 2025에 액세스, https://arxiv.org/html/2406.09039v2
134. Tutorial - RoboPoint VLM for Robotic Manipulation - NVIDIA Jetson AI Lab, 4월 25, 2025에 액세스, https://www.jetson-ai-lab.com/robopoint.html
135. 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/
136. Announcing new MoveIt2 Isaac SIM tutorial - MoveIt - ROS Discourse, 4월 25, 2025에 액세스, https://discourse.ros.org/t/announcing-new-moveit2-isaac-sim-tutorial/29991
137. 3D Mapping with an RGB-D Camera - Computer Vision Group - Technische Universität München, 4월 25, 2025에 액세스, https://cvg.cit.tum.de/_media/spezial/bib/endres2013tro.pdf
138. README.md - NVIDIA-AI-IOT/nanosam - GitHub, 4월 25, 2025에 액세스, https://github.com/NVIDIA-AI-IOT/nanosam/blob/main/README.md
139. End-to-End Intelligent Adaptive Grasping for Novel Objects Using an Assistive Robotic Manipulator - MDPI, 4월 25, 2025에 액세스, https://www.mdpi.com/2075-1702/13/4/275
140. Automatic Behavior Tree Expansion with LLMs for Robotic Manipulation - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.13356v1
141. moveit/moveit2: :robot: MoveIt for ROS 2 - GitHub, 4월 25, 2025에 액세스, https://github.com/moveit/moveit2
142. MoveIt 2 Binary Install, 4월 25, 2025에 액세스, https://moveit.ai/install-moveit2/binary/
143. Intent3D: 3D Object Detection in RGB-D Scans Based on Human Intention | OpenReview, 4월 25, 2025에 액세스, https://openreview.net/forum?id=5GgjiRzYp3
144. Language-driven Grasp Detection - CVF Open Access, 4월 25, 2025에 액세스, https://openaccess.thecvf.com/content/CVPR2024/papers/Vuong_Language-driven_Grasp_Detection_CVPR_2024_paper.pdf
145. Open-Vocabulary Part-Based Grasping - QUT ePrints, 4월 25, 2025에 액세스, https://eprints.qut.edu.au/253403/1/Tjeard%20Douwe%20van%20Oort%20Thesis%281%29.pdf
146. Tinker 2024 Team Description Paper - RoboCup@Home, 4월 25, 2025에 액세스, https://athome.robocup.org/wp-content/uploads/OPL-Tinker_2024_TDP.pdf
147. Grounding Dino + SAM2 for Image Segmentation with Text Inputs - YouTube,

4월 25, 2025에 액세스, https://www.youtube.com/watch?v=WCQ6XJ9wV1w
148. HiFi-CS: Towards Open Vocabulary Visual Grounding For Robotic Grasping Using Vision-Language Models - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/384074937_HiFi-CS_Towards_Open_Vocabulary_Visual_Grounding_For_Robotic_Grasping_Using_Vision-Language_Models
149. Robot Perception: Fine-Tuning YOLO with Grounded SAM 2 : r/computervision - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/computervision/comments/1hhe7n8/robot_perception_finetuning_yolo_with_grounded/
150. HiFi-CS: Towards Open Vocabulary Visual Grounding For Robotic Grasping Using Vision-Language Models - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2409.10419v1
151. Grounded-SAM-2/SAM2_README.md at main - GitHub, 4월 25, 2025에 액세스, https://github.com/IDEA-Research/Grounded-SAM-2/blob/main/SAM2_README.md
152. Robotic Arm Platform for Multi-View Image Acquisition and 3D Reconstruction in Minimally Invasive Surgery - UCL Discovery - University College London, 4월 25, 2025에 액세스, https://discovery.ucl.ac.uk/id/eprint/10206786/1/RA_L_formatted.pdf
153. pal-robotics/advanced_grasping_tutorials - GitHub, 4월 25, 2025에 액세스, https://github.com/pal-robotics/advanced_grasping_tutorials
154. RoboGrasp: A Universal Grasping Policy for Robust Robotic Control - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2502.03072v1
155. Plugin Interfaces - MoveIt, 4월 25, 2025에 액세스, https://moveit.ai/documentation/plugins/
156. Interactive Robotic Perception of Cable-Like Deformable Objects, 4월 25, 2025에 액세스, https://dspace.cvut.cz/bitstream/handle/10467/121584/F3-D-2025-Holesovsky-Ondrej-holesovsky-thesis.pdf?sequence=-1
157. ROS2 Manipulation Basics - Robotics & ROS Online Courses | The Construct, 4월 25, 2025에 액세스, https://app.theconstruct.ai/Course/81/
158. [Live Training] ROS2 Manipulation with Perception - The Construct, 4월 25, 2025에 액세스, https://www.theconstruct.ai/ros2-manipulation-training/
159. Simple ros2 grasp service : r/ROS - Reddit, 4월 25, 2025에 액세스, https://www.reddit.com/r/ROS/comments/1hhwo06/simple_ros2_grasp_service/
160. ROS 2 Manipulation Basics | The Construct, 4월 25, 2025에 액세스, https://www.theconstruct.ai/robotigniteacademy_learnros/ros-courses-library/ros2-manipulation-basics/
161. Pick and Place with the MoveIt Task Constructor for ROS 2 - Automatic Addison, 4월 25, 2025에 액세스, https://automaticaddison.com/pick-and-place-with-the-moveit-task-constructor-for-ros-2/
162. Welcome to ROS2 Grasp Library Tutorials, 4월 25, 2025에 액세스,

https://intel.github.io/ros2_grasp_library/
163. Robotic Grasping with 3D Visual Observations · robotology-legacy gym-ignition · Discussion #362 - GitHub, 4월 25, 2025에 액세스, https://github.com/robotology-legacy/gym-ignition/discussions/362
164. NVIDIA Brings Generative AI Tools, Simulation and Perception Workflows to ROS Developer Ecosystem, 4월 25, 2025에 액세스, https://blogs.nvidia.com/blog/generative-ai-simulation-roscon/
165. ROS2 NanoOWL for Open-vocabulary object detection - NVIDIA Developer Forums, 4월 25, 2025에 액세스, https://forums.developer.nvidia.com/t/ros2-nanoowl-for-open-vocabulary-object-detection/291296
166. Multiple Robot ROS2 Navigation - Isaac Sim Documentation, 4월 25, 2025에 액세스, https://docs.isaacsim.omniverse.nvidia.com/4.5.0/ros2_tutorials/tutorial_ros2_multi_navigation.html
167. Robot Manipulation with ROS2 using MoveIt2 Training - One-Day Online Training, 4월 25, 2025에 액세스, https://www.theconstruct.ai/one-day-training/robot-manipulation-moveit2/
168. MorphoNavi: Aerial-Ground Robot Navigation with Object Oriented Mapping in Digital Twin - arXiv, 4월 25, 2025에 액세스, https://arxiv.org/html/2504.16914v1
169. Playful DoggyBot: Learning Agile and Precise Quadrupedal Locomotion - ResearchGate, 4월 25, 2025에 액세스, https://www.researchgate.net/publication/384501688_Playful_DoggyBot_Learning_Agile_and_Precise_Quadrupedal_Locomotion
170. [ROS Courses] Mastering with ROS: Smart Grasping Systml | Robot Ignite Academy, 4월 25, 2025에 액세스, https://www.theconstruct.ai/robotigniteacademy_learnros/ros-courses-library/mastering-ros-smart-grasping-system/
171. Robot multi-target high performance grasping detection based on random sub-path fusion, 4월 25, 2025에 액세스, https://pmc.ncbi.nlm.nih.gov/articles/PMC11906837/
172. ICRA 2025 Program | Wednesday May 21, 2025, 4월 25, 2025에 액세스, https://ras.papercept.net/conferences/conferences/ICRA25/program/ICRA25_ContentListWeb_2.html