단일 RGB 이미지 기반 3D HOI 복원을 위한 오픈보캐뷸러리 어포던스 기반 적응형 특징 융합 및 최적화 가이드 기법

우승우, 김지원, 염세경
동국대학교

woow0708@dgu.ac.kr, kimji1net@dgu.ac.kr, skyoum@dongguk.edu

Adaptive Feature Fusion and Optimization Guidance based on Open-Vocabulary Affordance for 3D HOI Reconstruction from a Single RGB Image

Woo Seung Woo, Ji Won Kim, Youm Se Kyoung
Dongguk Univ.

요 약

본 논문은 단일 RGB 이미지에서 인체 모델과 물체 메시 간의 정밀 접촉면을 복원하기 위해 오픈보캐뷸러리, OV 어포던스 지식과 기하학 기반 접촉 예측을 결합하는 적응형 특징 융합 프레임워크를 제안한다. 기존 최적화 기반 HOI, Human Object Interaction 복원은 단일 시점의 깊이 모호성과 가림 현상으로 인해 기능적으로 부적절한 부위에 접촉이 수렴해 고착되는 문제가 빈번하다. 이를 완화하기 위해 본 연구는 메시 인코더가 출력하는 기하학적 접촉 확률과 비전 언어 모델, VLM 기반 프롬프트로부터 추출된 어포던스 맵을 게이팅과 가중 결합으로 융합해 정제된 접촉 가이드를 생성한다. 또한 정제된 가이드를 최적화 루프에 주입할 때 잔차 형태와 stop gradient 를 적용해 물리적 제약을 훼손하지 않으면서 의미론적 제약을 안정적으로 반영한다. 제안 기법은 미지 물체에 대해서도 기능적으로 타당한 접촉면 형성을 유도함으로써 기하 중심 최적화 방식의 강건성을 향상시키는 것을 목표로 한다.

Ⅰ. 서 론

단일 RGB 이미지 기반 3 차원 인체 물체 상호작용 복원은 가상현실, 디지털 휴먼, 로봇 조작과 내비게이션 등 다양한 응용의 핵심 기술이다. 그중에서도 정밀 접촉면 복원은 상호작용의 물리적 타당성을 좌우하는 핵심 요소이나, 단일 시점에서는 깊이 모호성과 가림으로 인해 접촉 지점을 안정적으로 식별하기 어렵다.

SMPL-X 와 같은 파라메트릭 인체 모델을 활용하는 접근은 대체로 기하학적 근접도와 비관통 제약에 의존해 접촉을 유도해 왔다[1]. 그러나 단일 시점 최적화에서는 초기화의 불완전성과 관측의 제약으로 인해, 기하학적으로는 근접하지만 기능적으로는 부적절한 표면에 접촉이 수렴하는 문제가 발생할 수 있다. 예를 들어 잡기 상호작용에서 칼날이나 용기 외벽과 같이 기능적으로 부적절한 영역이 접촉 후보로 고정되는 현상이 나타난다. 최근 PICO-fit[7]은 이미지 기반 적합 과정에서 접촉 제약을 활용해 HOI 복원을 개선했으나, 단일 시점에서의 모호성이 큰 경우 접촉의 의미론적 타당성을 보장하기 어렵다는 한계가 남는다.

본 논문은 이러한 한계를 완화하기 위해 오픈보캐뷸러리 어포던스 지역화 기술인 OVAL-Prompt[4]와 OOAL[5]의 흐름을 접촉 복원에 통합하여, 텍스트로 정의된 상호작용 맥락을 접촉면 추론 과정에 직접 주입하는 파이프라인을 제안한다. 제안 방법은 첫째, VLM 프롬프트로부터 도출되는 어포던스 맵을 접촉 복원에 직접 활용해 기능적으로 타당한 후보 영역을 제공하고, 둘째, 기하학 접촉 확률과 어포던스 맵을 게이팅과 가중 결합으로 통합해 오탐 접촉을 억제하며, 셋째, 잔차 주입과 stop gradient 를 적용해 의미론 신호가 물리적 제약을 과도하게 왜곡하지 않도록 최적화 루프의 안정성을 확보한다.

Ⅱ. 본론

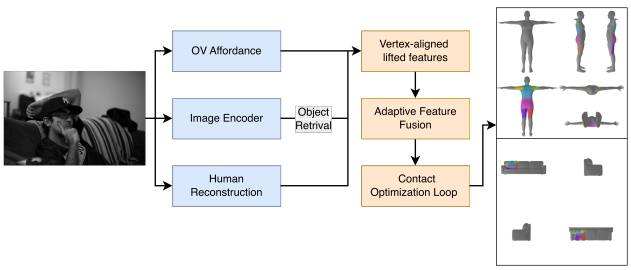

Figure 1: System Frameworks

본 논문에서 제시한 시스템은 단일 RGB 이미지를 입력으로 받아 인체와 물체의 초기 3 차원 메시를 복원한

뒤, 이미지 기반 시각 특징과 정점 단위 기하 정보를 결합해 접촉 필드를 추론하고, 프롬프트 기반 지역화 결과를 접촉 가이드로 주입해 최적화 과정에서 정밀 접촉면을 유도하는 단계로 구성된다. 효율적인 학습과 안정적인 시각 지식 활용을 위해 이미지 특징 추출에는 DINOv2[3]를 사용하고, 인체 복원 네트워크는 SMPLest-X[2] 계열의 사전 학습 가중치를 고정해 초기화를 제공하는 구성으로 기술한다.

시스템의 시작점은 입력 이미지로부터 인체와 물체의 초기 메시를 획득하는 것이다. 인체는 SMPLest-X[2]를 통해 SMPL-X 파라미터를 추정하여 메시를 복원하고, 물체는 DINOv2 특징 기반 유사도 검색을 활용해 Objaverse[6]로부터 유사 메시를 선택하여 초기화한다. 이후 각 메시 정점의 3 차원 좌표를 카메라 모델로 이미지 평면에 투영하고, 투영 위치에서 이미지 특징을 샘플링하여 정점에 할당한다. 이로써 각 정점은 3 차원 기하 정보와 이미지 기반 시각 특징이 결합된 정점 정렬 특징을 갖게 되며, 이는 후속 메시 인코더의 입력으로 사용된다.

본 연구의 핵심은 프롬프트 유도형 어포던스 지역화 모듈의 도입이다. 시스템은 잡을 수 있는 부위, 받칠 수 있는 바닥과 같은 기능 중심 텍스트 질의를 입력받아 이미지 내에서 해당 기능을 수행하는 영역을 지역화하고, 이를 어포던스 맵 형태로 정규화해 출력한다. 이 어포던스 맵은 물체의 외형을 넘어 상호작용이 가능한 의미론적 영역을 지정하는 제약으로 작동하므로, 단일 시점 접촉 복원에서 발생하는 기능적으로 부적절한 접촉 수렴을 억제하는 역할을 수행한다.

이후 메시 인코더가 출력하는 기하학적 접촉 확률과 어포던스 맵을 적응적으로 융합하여 정제된 접촉 가이드를 생성한다. 융합 단계에서는 어포던스 맵이 시맨틱 필터로 작동해, 기하학적으로는 접촉 가능성이 높게 예측되더라도 텍스트로 정의된 기능 영역에서 벗어난 정점의 접촉 확률을 억제한다. 또한 두 신호를 단순 합산하지 않고, 게이팅과 가중 결합으로 결합해 어포던스 신뢰도가 낮은 영역에서의 과도한 억제를 방지하고, 기하학 신호의 장점을 유지하도록 설계한다.

정제된 접촉 가이드는 PICO-fit[7]과 같은 최적화 루프에 주입되어 인체가 물체의 사용 가능한 부위로 수렴하도록 유도한다. 이때 어포던스 사전 지식의 주입 과정에는 잔차 형태와 stop gradient 를 포함한 주입 규칙을 적용하여, 의미론적 가이드가 물리적 접촉 신호의 고유한 특성을 과도하게 변형하지 않도록 한다.

정밀한 접촉 복원과 물리적 타당성을 위해 학습 목적 함수는 세 가지 축으로 구성한다. 첫째 접촉 필드 감독 손실은 최종 융합 접촉 확률이 정답 접촉과 일치하도록 학습하며 접촉 희소성을 고려해 포컬 손실을 적용한다[8]. 둘째 어포던스 일관성 손실은 예측 접촉이 어포던스 맵의 고신뢰 영역과 정합되도록 유도하여 기능적으로 부적절한 접촉을 억제한다. 셋째 기하학 제약 손실은 비관통 손실과 메시 평활화 정규화를 결합하여 인체 정점의 물체 내부 침투를 방지하고 접촉 지도의 공간적 연속성을 확보한다. 결과적으로 본 시스템은 프롬프트 기반 어포던스 지역화와 적응형 특징 융합을 통해 의미론 정보가 내재된 접촉 필드를 생성하고, 기하학 근접성만으로 발생하던 의미적 부적절성을 완화한 HOI 복원을 달성하는 것을 목표로 한다.

Ⅲ. 결론

본 연구는 단일 이미지 기반 3D HOI 복원에서 접촉이 기능적으로 부적절한 부위에 고착되는 문제를 완화하기 위해, 프롬프트 기반 오픈보캐뷸러리 어포던스 지역화 결과를 접촉 복원 최적화 과정에 통합하는 설계안을 제시하였다. 기하학 접촉 확률과 어포던스 맵을 게이팅과 가중 결합으로 적응적으로 융합하고, 잔차 주입과 stop gradient 기반 규칙으로 최적화 루프에 안정적으로 주입함으로써 물리적 타당성과 의미론적 타당성을 동시에 확보하고자 하였다. 향후 연구에서는 실제 HOI 데이터셋을 활용한 정량 평가를 수행하고, 프롬프트 조합 변화에 따른 일반화 성능과 실패 사례를 체계적으로 분석할 계획이다.

참 고 문 헌

[1] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, “Expressive Body Capture: 3D Hands, Face, and Body from a Single Image,” Proc. IEEE CVF Conf. on Computer Vision and Pattern Recognition, 2019, pp. 10975- 10985.

[2] W. Yin et al., “SMPLest-X: Ultimate Scaling for Expressive Human Pose and Shape Estimation,” arXiv:2501.09782, 2025.

[3] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” arXiv:2304.07193, 2023.

[4] E. Tong, A. Opipari, S. Lewis, Z. Zeng, and O. C. Jenkins, “OVAL-Prompt: Open-Vocabulary Affordance Localization for Robot Manipulation through LLM Affordance-Grounding,” arXiv:2404.11000, 2024.

[5] G. Li, D. Sun, L. Sevilla-Lara, and V. Jampani, “One-Shot Open Affordance Learning with Foundation Models,” Proc. IEEE CVF Conf. on Computer Vision and Pattern Recognition, 2024.

[6] M. Deitke et al., “Objaverse: A Universe of Annotated 3D Objects,” Proc. IEEE CVF Conf. on Computer Vision and Pattern Recognition, 2023.

[7] A. Cseke et al., “PICO: Reconstructing 3D People In Contact with Objects,” Proc. IEEE CVF Conf. on Computer Vision and Pattern Recognition, 2025.

[8] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” Proc. IEEE Int. Conf. on Computer Vision, 2017, pp. 2980- 2988.