서강대학교
SOGANG UNIVERSITY

2026 하계 세미나

Training-Free 6D object pose estimation

2026. 07. 10.

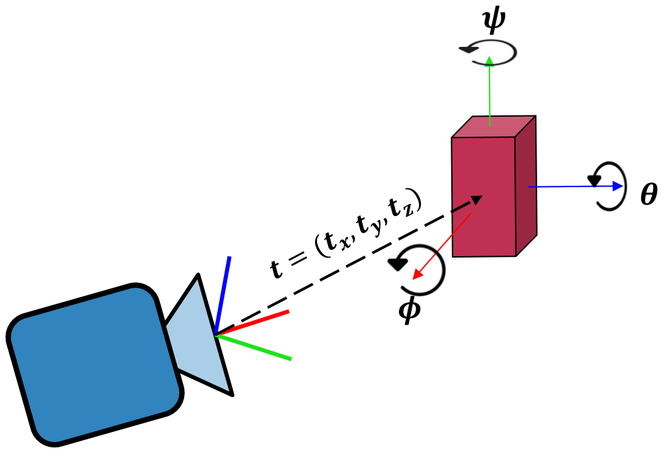
Presented By
김현빈

Outline

- Introduction
  - 6D object pose estimation
  - Unseen Object Pose Estimation
  - Foundation Features
- Paper Review
  - FoundPose: Unseen Object Pose Estimation with Foundation Features (ECCV 2024)
  - Pos3R: 6D Pose Estimation for Unseen Objects Made Easy (CVPR 2025)

Introduction

- 6D object pose estimation
  - 단일 RGB 이미지로부터 물체의 3D translation과 3D rotation을 추정하는 문제
    - **6D pose**: 3D translation (x, y, z) + 3D rotation (roll, pitch, yaw)
    - **Input**: RGB image or RGB-D image
    - **Output**: Object pose (Rotation, translation) in **camera coordinate system**
  - 6D pose in Real-World Application
    - Robotic Manipulation, AR / VR



< 6D pose representation of a target object >

Estimated Object Poses AR Demo

< Pose estimation for AR application >

< Pose estimation for robotic manipulation >

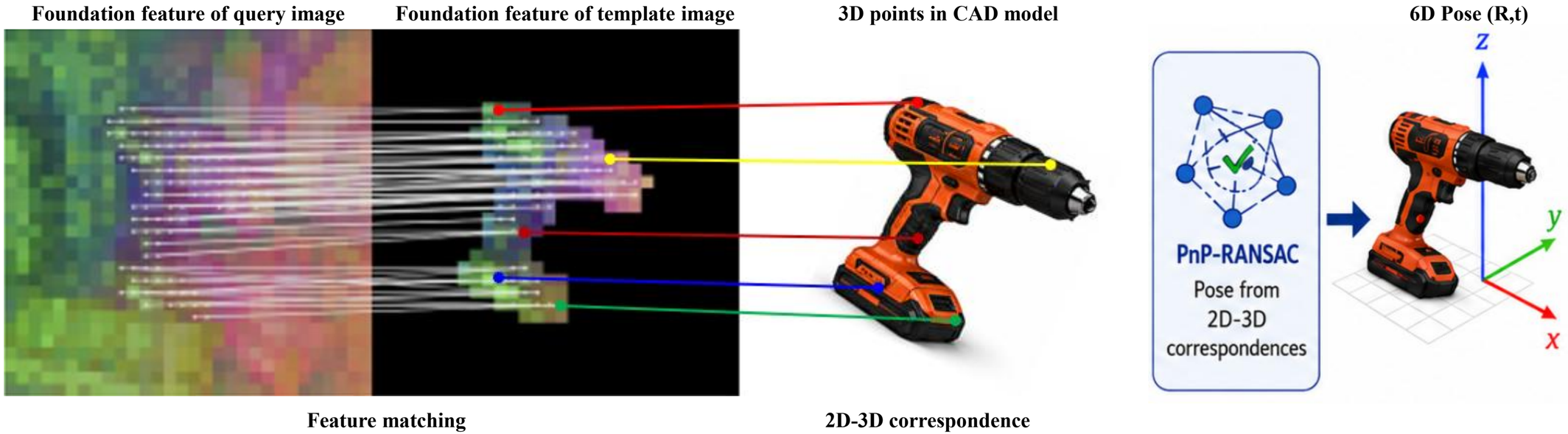
Introduction

- Unseen Object Pose Estimation
  - Unseen Object 문제
    - 학습 시점에 존재하지 않던 새로운 물체의 pose를 추정해야 함
    - 산업·물류 현장에서는 대상 물체가 수시로 바뀌어 물체별 재학습이 비현실적
  - 기존 학습 기반 방법의 한계
    - 물체 instance마다 대규모 annotation과 재학습 필요 → 확장성 저하
    - 최근 unseen 방법도 대규모 학습에 의존
      - MegaPose·GigaPose·GenFlow: 수천 개 물체의 수백만 장 합성 데이터로 render-and-compare 학습
      - 데이터 생성 비용·물체 texturing·scene 배치 등 새로운 부담 발생

→ **따라서 학습 없이 즉시 onboarding 가능한 training-free 방법이 요구됨**

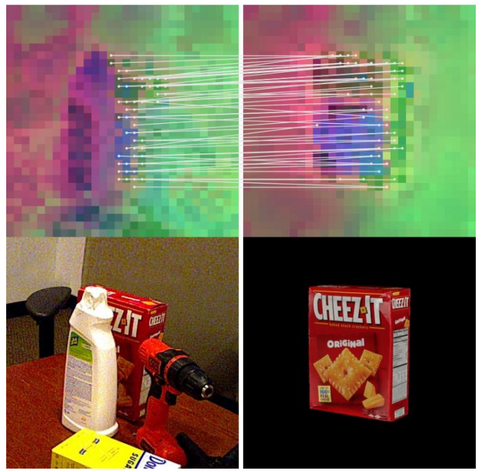
Introduction

- Foundation Features
  - Foundation feature는 pose를 직접 예측하지 않고 real image와 CAD-rendered template 사이의 correspondence를 제공
    - 추출된 feature 유사도를 비교하여 query patch와 template patch를 매칭
    - Template의 depth 또는 coordinate map을 이용해 2D-3D correspondence 구성
    - 이후 PnP-RANSAC과 refinement가 rotation R과 translation t를 기하적으로 계산

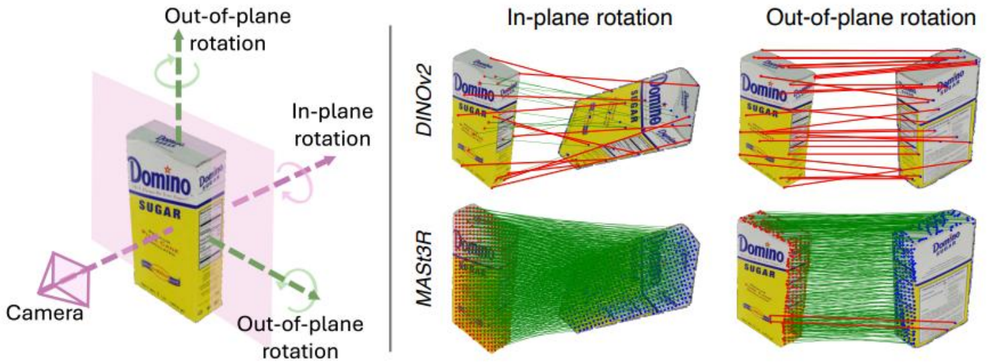


Introduction

- Foundation Features
  - 2D Foundation Model: DINOv2 → FoundPose
    - 중간 layer의 patch descriptor는 별도 학습 없이 실제 이미지와 합성 이미지 간 correspondence를 형성함
    - 2D appearance 기반 feature이므로 out-of-plane viewpoint 변화에 취약함
  - 3D Foundation Model: MASt3R → Pos3R
    - 3D reconstruction + pixel-wise matching 공동 학습 → 시점 간 3D-consistent feature
    - 서로 다른 viewpoint에서도 같은 3D surface를 안정적으로 연결



< DINOv2-based Synthetic-to-Real Correspondence >



< DINOv2 vs. MASt3R correspondence under viewpoint changes >

1) Örnek, Evin Pınar, et al. "Foundpose: Unseen object pose estimation with foundation features." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

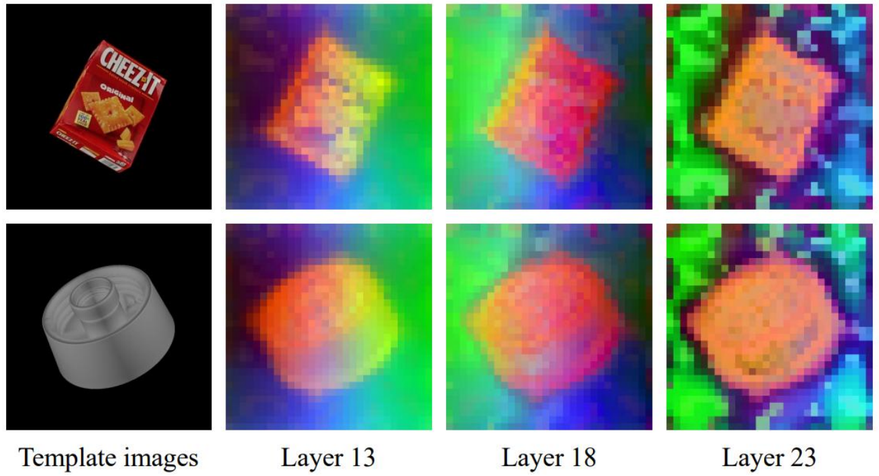
Paper Review

FoundPose: Unseen Object Pose Estimation with Foundation Features (ECCV 2024)

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, Tomas Hodan

Introduction

- Motivation
  - 기존 training-free 방법은 synthetic-to-real domain gap 해소에 대규모 학습이 필요
    - MegaPose: 2M+ 합성 이미지 학습 / OSOP: 물체당 9만+ template
    - 따라서 학습 없이 domain gap을 넘는 방법 필요



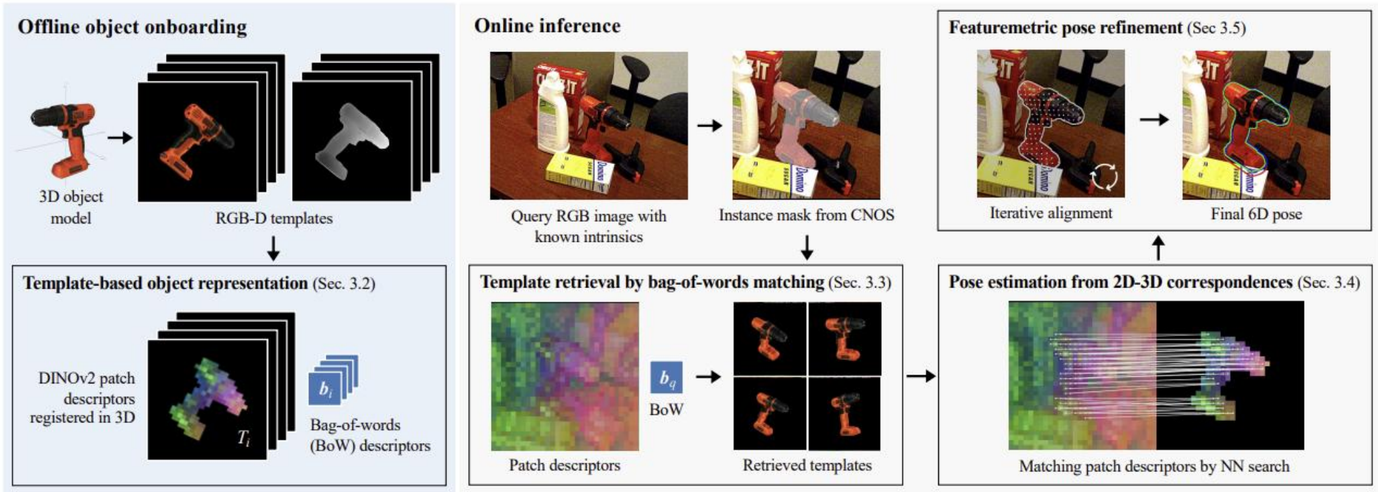
Template images | Layer 13 | Layer 18 | Layer 23

< Visualization of DINOv2 patch descriptors >

  - DINOv2 patch descriptor에서 pose estimation에 필요한 특성을 관찰
    - 얕은 layer일수록 물체의 면·위치를 구분하는 위치 정보가 강함
    - 깊은 layer일수록 부위 의미를 담는 semantic 정보 위주로 추출
    - 중간 layer가 두 정보를 고루 추출할 수 있음
      - Texture-less 물체는 semantic이 모호해지나, positional 정보가 같은 patch끼리 매칭을 유도할 수 있음
        → 기하학적으로 일관된 correspondence을 확보할 수 있음
  - FoundPose
    - Foundation feature를 고전 방법(BoW + PnP)에 결합한 training-free pipeline을 제안
    - 추가 학습 없이 BOP benchmark RGB SOTA 달성

Methodology

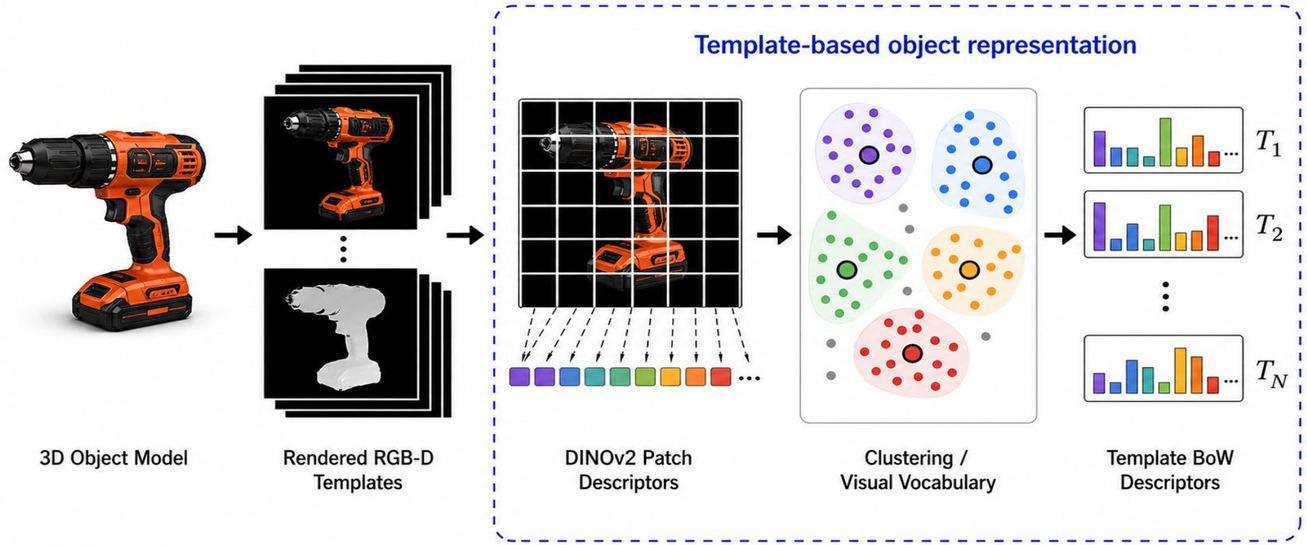
- Overview
  - Offline onboarding
    - 3D model에서 여러 viewpoint의 RGB-D template 렌더링 후, 각 template에서 DINOv2 patch descriptor 추출
    - Patch descriptor를 depth 정보를 이용해 3D model point와 연결
  - Online inference
    - Query image에서 object crop을 생성한 뒤, 유사한 template을 검색
    - Query crop과 template 간 patch matching 수행 후, 2D-3D correspondence를 이용해 PnP-RANSAC으로 pose 추정
  - Key Idea
    - **DINOv2 feature를 이용해 real image와 synthetic template 사이의 correspondence를 학습 없이 형성**



< Pipeline of FoundPose >

Methodology

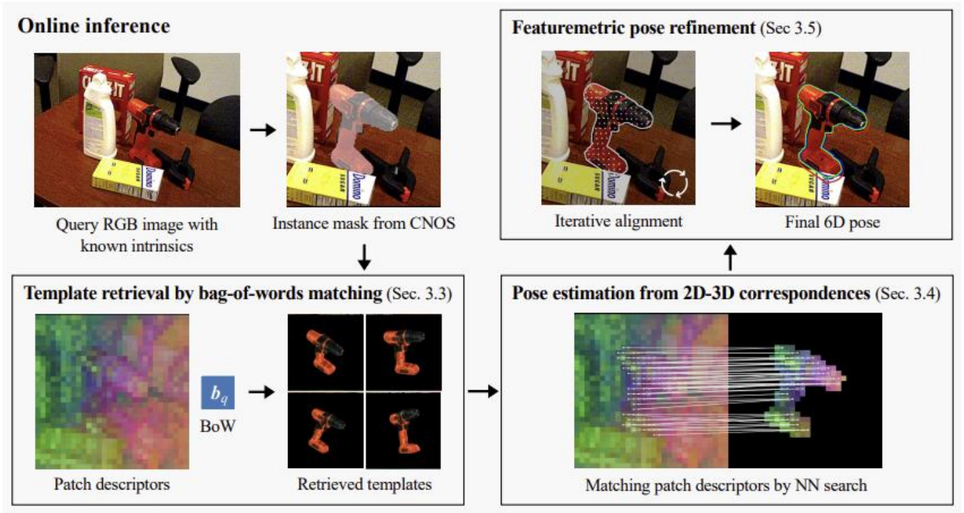
- Offline onboarding: Template-based object representation
  - Pipeline
    1. CAD 모델을 여러 viewpoint에서 렌더링하여 template image 생성
    2. 각 template에서 DINOv2 patch descriptor 추출 → 3D model point와 연결
       - Query-template patch matching 시 2D-3D correspondence로 사용
    3. PCA로 descriptor 차원 축소 후, K-means clustering으로 visual vocabulary 생성
    4. Template마다 visual word histogram을 구성해 Bag-of-Words(BoW) descriptor로 저장



< Template-based object representation >

Methodology

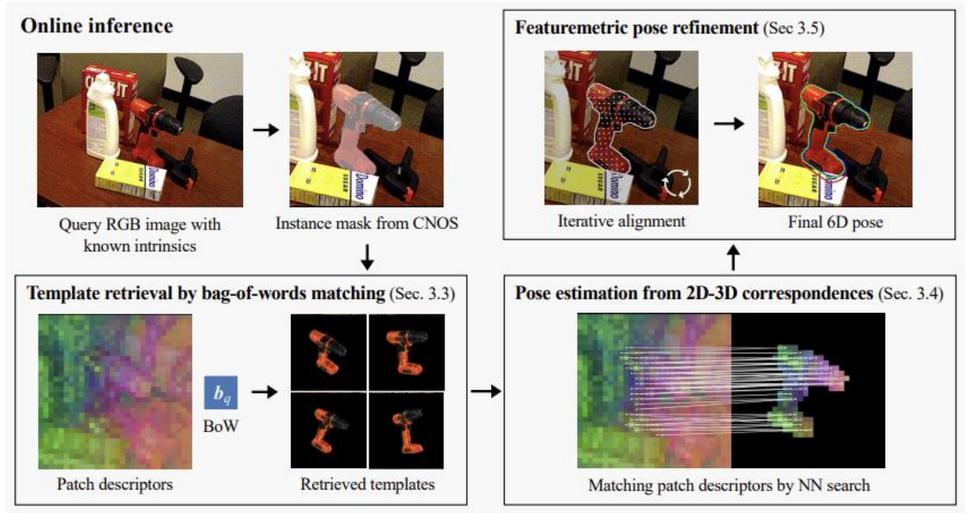
- Online inference: Template retrieval by bag-of-words matching
  - Pipeline
    - 입력 RGB image에서 object mask(CNOS)를 추출하고, perspective crop을 생성
      - 원근 왜곡을 줄여 template과 유사한 형태의 crop을 얻음
    - Crop에서 DINOv2 중간 layer patch descriptor를 추출
    - Offline과 동일한 PCA projection과 K-means visual vocabulary를 적용해 query patch를 visual word로 양자화
    - Template BoW descriptor와 비교하여 유사한 top-K template를 검색 → Pose 후보로 선정



< Pipeline of Online inference >

Methodology

- Online inference: Pose estimation from 2D–3D correspondences
  - Pose Estimation from 2D–3D Correspondences
    - Retrieved template의 patch를 query patch와 kNN으로 매칭
    - Template patch에 등록된 CAD surface의 3D 좌표를 query patch의 3D 대응점으로 사용
      - Onboarding 때 depth로 각 patch에 3D 좌표를 미리 등록해 둠
    - Query 2D 위치 ↔ CAD 3D point로 2D–3D correspondence set을 구성 (retrieved template마다 하나씩)
    - 각 correspondence set에서 EPnP + RANSAC으로 pose를 추정
      - RANSAC inlier가 가장 많은 pose를 최종 coarse pose로 선택



< Pipeline of Online inference >

Methodology

- Online inference: Featuremetric Pose Refinement
  - Coarse sampling으로 인한 2D-3D 불일치
    - Patch가 14×14px로 크기 때문에, 대응점이 patch 단위로만 샘플링됨
    - GT pose로 투영해도 patch 중심과 실제 3D projection이 어긋남 → PnP 입력 자체에 양자화 오차가 내재
    - 이 오차를 discrete correspondence가 아닌 연속적인 pose 공간에서 직접 보정
  - Featuremetric Alignment
    - Photometric alignment(Lucas-Kanade)을 DINOv2 feature map에 적용
      - 매 스텝마다 선형회귀하듯이 R, t를 맞춰나가는 방식

$$(\mathbf{R}_r, \mathbf{t}_r) = \underset{(\mathbf{R},\mathbf{t})}{\operatorname{argmin}} \sum_{(\mathbf{p}_i, \mathbf{x}_i) \in T_t} \rho\left(\mathbf{p}_i - \mathbf{F}_q\left(\pi(\mathbf{R}\mathbf{x}_i + \mathbf{t})/s\right)\right),$$

      - p_i, x_i : template patch의 descriptor와 대응 3D 좌표
      - $\mathbf{F}_q$: query feature map, $\pi(\cdot)/s$: 3D점을 2D로 투영 후 patch 크기 s로
      - ρ: Barron robust loss ($\alpha = -5,\ c = 0.5$) → outlier 대응점 억제

Experiments

- Dataset
  - BOP 7개 core dataset (LM-O·T-LESS·TUD-L·IC-BIN·ITODD·HB·YCB-V)
- Metric
  - BOP Average Recall (AR):
    - AR = (AR_VSD + AR_MSSD + AR_MSPD) / 3, 세 pose error 지표의 평균
      - VSD (Visible Surface Discrepancy)
        - ✓추정 pose와 GT pose로 각각 물체를 렌더링해 얻은 depth map을 비교
      - MSSD (Max **Symmetry-aware** Surface Distance)
        - ✓모델 꼭짓점들의 3D 거리의 최댓값 (단위: 객체 크기 대비 비율) – 가장 많이 차이나는 꼭짓점 탐색
        - ✓대칭 고려한 3D 표면 거리 오차
      - MSPD (Max **Symmetry-aware** Projection Distance)
        - ✓표면 점들을 이미지 평면에 투영한 뒤, 추정 pose와 GT pose 간 2D 픽셀 거리의 최댓값을 측정(단위: 픽셀)
        - ✓대칭 고려한 2D 투영 오차
      - 각 지표는 여러 correctness 임계값에서 맞았는지 여부를 판정하고, 맞은 비율(recall)을 임계값에 대해 평균

Experiments

- Comparison with SOTA
  - Coarse pose estimation
    - 학습 기반 방법(GigaPose, GenFlow, MegaPose) 대비 평균 AR이 높음
    - 추론 시간은 1.7s로, render-and-compare 계열(MegaPose 15.5s)보다 짧음
  - With pose refinement
    - Featuremetric only: 학습 없이 coarse 대비 평균 +5 AR
    - MegaPose refiner 결합 시 59.6 AR로, RGB-only 중 가장 높음

#	Method	Pose refinement	No train.	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Average	Time
Coarse pose estimation:												
1	FoundPose	–	✓	**39.6**	**33.8**	**46.7**	**23.9**	**20.4**	**50.8**	**45.2**	**37.2**	1.7
2	GigaPose [59]	–	✗	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6	**0.9**
3	GenFlow [54]	–	✗	25.0	21.5	30.0	16.8	15.4	28.3	27.7	23.5	3.8
4	MegaPose [41]	–	✗	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8	15.5
5	OSOP [75]	–	✗	31.2	–	–	–	–	49.2	33.2	–	–
6	ZS6D [3]	–	✓	29.8	21.0	–	–	–	–	32.4	–	–
With pose refinement (5 hypotheses):												
12	FoundPose	Featuremetric	✓	42.0	43.6	60.2	30.5	27.3	53.7	51.3	44.1	7.4
13	FoundPose	MegaPose	✗	58.6	54.9	65.7	44.4	36.1	70.3	67.3	56.8	11.2
14	FoundPose	Feat. + MegaPose	✗	**61.0**	**57.0**	**69.4**	**47.9**	**40.7**	72.3	**69.0**	**59.6**	20.5
15	GigaPose [59]	MegaPose	✗	59.9	**57.0**	64.5	46.7	39.7	72.2	66.3	57.9	**7.3**
16	GenFlow [54]	GenFlow	✗	56.3	52.3	68.4	45.3	39.5	**73.9**	63.3	57.1	20.9
17	MegaPose [41]	MegaPose	✗	56.0	50.7	68.4	41.4	33.8	70.4	62.1	54.7	47.4

< Comparison with SOTA >

Experiments

- Ablation study
  - Intermediate Layer
    - 동일 backbone(DINOv2 ViT-L)에서 중간 layer 18 = 37.2 AR, 마지막 layer 23 = 23.5 AR
    - ViT-S에서도 같은 경향을 보임 (layer 9 = 34.0 vs layer 11 = 23.7)
  - Feature Extractor
    - DINOv2 외 다른 모델로 교체 시 정확도가 크게 하락: SAM 10.7 / DenseSIFT 7.6 / S2DNet 1.1 AR
  - Effect of mask
    - GT mask로 교체 시 +6–19 AR (rows 1 vs 11) → mask 품질이 주요 요인으로 작용할 수 있음

#	Method	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Average	Time
	Backbones for extracting patch descriptors:									
1	DINOv2 ViT-L – layer 18	39.6	33.8	46.7	23.9	20.4	50.8	45.2	37.2	1.7
2	DINOv2 ViT-L – layer 23	23.2	22.8	31.2	10.3	9.7	33.0	34.0	23.5	1.5
3	DINOv2 ViT-S – layer 9	34.0	31.6	42.7	21.7	16.8	46.8	44.7	34.0	1.3
4	DINOv2 ViT-S – layer 11	22.8	24.2	29.8	11.9	10.5	30.4	36.4	23.7	1.3
5	SAM ViT-L [40] – layer 23	2.2	12.8	9.2	7.5	6.0	10.6	26.9	10.7	3.4
6	DenseSIFT – step size 7px	3.2	2.6	6.5	10.5	2.9	5.6	22.2	7.6	1.4
7	S2DNet [23]	0.8	1.2	0.8	1.4	1.2	1.2	1.3	1.1	1.8
10	Pose given by the top matched template	20.3	18.5	23.0	12.8	12.4	19.6	17.6	17.7	1.0
11	Ground-truth instead of CNOS masks	45.6	53.1	57.1	30.6	–	–	50.9	–	–

< Ablation experiments >

1) Deng, Weijian, et al. "Pos3r: 6d pose estimation for unseen objects made easy." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

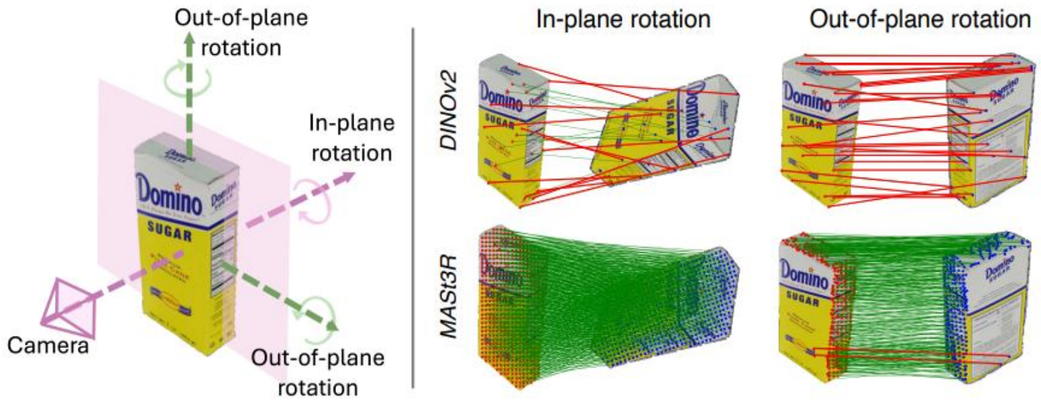
Paper Review

Pos3R: 6D Pose Estimation for Unseen Objects Made Easy (CVPR 2025)

Weijian Deng, Dylan Campbell, Chunyi Sun, Jiahao Zhang, Shubham Kanitkar, Matthew E. Shaffer, Stephen Gould

Introduction

- Motivation
  - 2D Foundation Feature의 한계
    - In-plane rotation에서는 DINOv2가 비교적 잘 작동
    - Out-of-plane rotation은 시점 변화로 appearance가 크게 변함
      - 그 결과 correspondence가 희소하고 불안정해질 수 있음



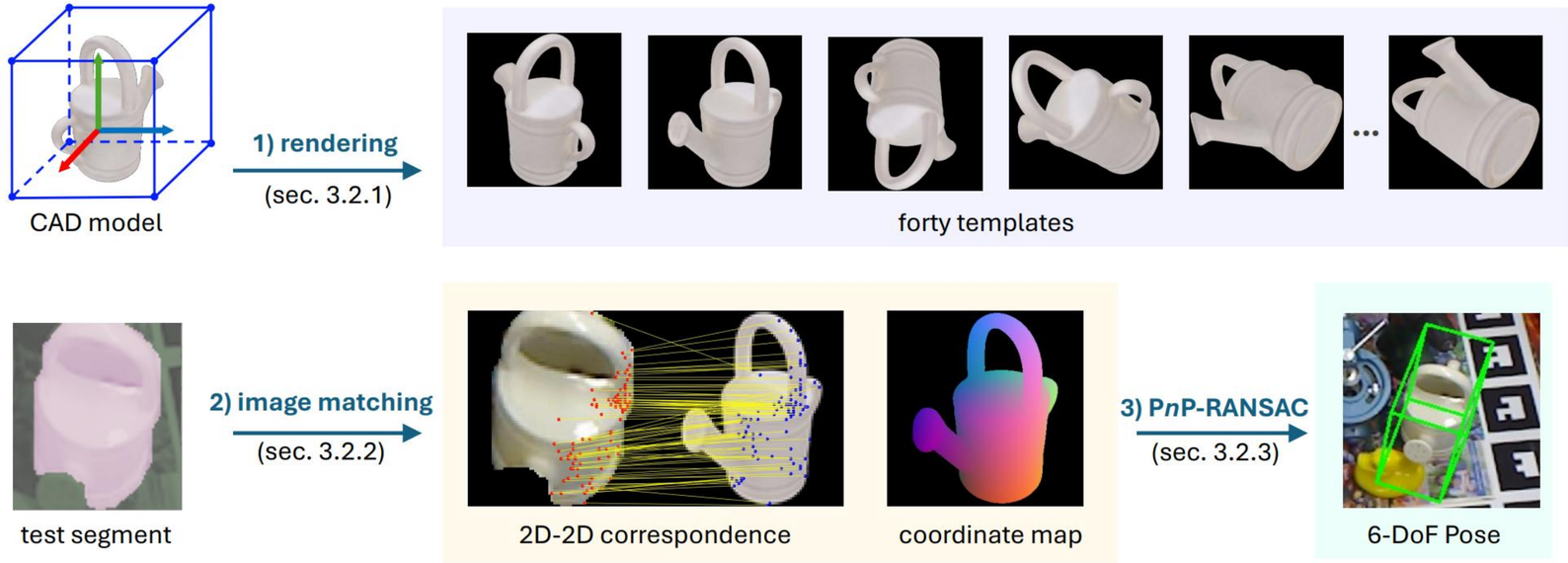
< DINOv2 vs. MASt3R under viewpoint changes >

  - 3D Foundation Model로 Viewpoint Gap 완화
    - MASt3R는 image pair의 3D structure를 함께 추정하며 pixel-wise matching 수행
      - 단순한 2D appearance similarity가 아니라, 서로 다른 시점에서 관측된 같은 3D surface를 연결
    - 따라서 out-of-plane rotation처럼 object projection이 크게 달라지는 경우에도 안정적인 matching 가능

  - Pos3R
    - MASt3R 기반 training-free 6D pose estimation
    - 40 templates만으로 template selection 수행
      - FoundPose 대비 template 수를 크게 절감(≈800 templates)

Methodology

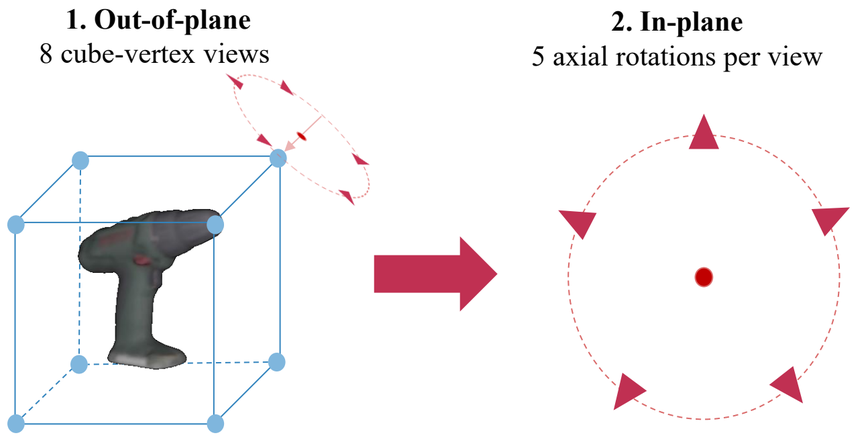
- Overall pipeline
  - Template Rendering
    - CAD model에서 40개 RGB template + coordinate map 생성
  - Image Matching
    - MASt3R를 활용하여 query segment와 template 간 correspondence 계산 → Matching quality가 가장 높은 template 선택
  - Pose Fitting
    - 선택된 template에 대응하는 coordinate map으로 2D-3D correspondence를 만들고 PnP-RANSAC 수행



< Pipeline of Pos3R >

Methodology

- Template Rendering & Configuration
  - Template Rendering
    - 8개 cube vertex view로 out-of-plane rotation cover
    - 각 view에 5개 axial rotation을 적용해 in-plane rotation 보완
    - 최종적으로 8 × 5 = 40 templates 구성



8 views × 5 in-plane rotations = 40 templates

Out-of-plane coverage from cube vertices,
in-plane coverage from axial rotations

  - Coordinate map
    - 각 template pixel에 대응하는 3D model coordinate 저장
    - 이후 matched template pixel을 3D point로 변환하는 데 사용

  - Why only 40 templates?
    - MASt3R feature가 viewpoint 변화에 강해 수백 개 template이나 selection network가 필요하지 않음

Methodology

- MASt3R Matching & Pose Fitting
  - Dense image matching
    - Query segment I_m와 template $I_{i,k}$를 MASt3R에 입력하여 픽셀별 feature 생성

$$f(I_m, I_{i.k}) = Dec\left(Enc(I_m), Enc\left(I_{i,k}\right)\right)$$

    - Feature map에서 reciprocal matching으로 대응 픽셀 좌표쌍을 생성

$$M_{m,i,k} = \{(y_m^p, y_{i,k}^p)\}$$

  - Template selection
    - 각 대응쌍의 feature similarity를 누적해 template score를 계산, 최고 점수 template 선택

$$sim\left(I_m, I_{i,k}\right) = p\sum_p f_m^p \cdot f_{i,k}^p, \qquad (i*, k*) = argmax_{i,k}\, sim(I_m, I_{i,k})$$

  - Pose fitting
    - 선택된 template의 매칭 픽셀을 coordinate map으로 3D 점에 대응시켜 2D–3D correspondence 구성
    - PnP-RANSAC으로 reprojection error를 최소화하는 pose 추정

$$argmin_{R,t} \sum_j \left\| y_m^j - \pi(RP^j + t) \right\|^2$$

Experiments

- Comparison with SOTA
  - Coarse pose estimation
    - Training-free 방법 중 가장 높은 성능을 보임(mean AR 39.5)
      - TUD-L·IC-BIN·ITODD·HB·YCB-V에서 최고, LM-O·T-LESS에서는 FoundPose가 우위
        - ✓약세 원인은 두 데이터셋이 서로 다름
          - LM-O: 심한 occlusion이 초기 matching을 방해하기 때문
          - T-LESS: 강한 대칭·물체 유사성으로 template selection이 어려움
    - 추론 시간은 1.4s로, render-and-compare 계열보다 짧음
      - FoundPose(1.7s, 798 templates)보다도 짧으며, template 수 감소가 요인으로 추정

#	Method	Training-Free	Refinement	Datasets (num. instances)							Mean	Time
				LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V		
				1445	6423	600	1786	3041	1630	4123		
Coarse Pose Estimation:												
1	GigaPose [39]	✗	–	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6	0.9
2	GenFlow [34]	✗	–	25.0	21.5	30.0	16.8	15.4	28.3	27.7	23.5	3.8
3	MegaPose [22]	✗	–	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8	15.5
4	OSOP [48]	✗	–	31.2	–	–	–	–	49.2	33.2	–	–
5	ZS6D [2]	✓	–	29.8	21.0	–	–	–	–	32.4	–	–
6	FoundPose [41]	✓	–	**39.6**	**33.8**	46.7	23.9	20.4	50.8	45.2	37.2	1.7
7	Baseline	✓	–	30.5	23.6	43.2	21.4	15.4	42.5	49.3	32.3	0.7
7	Pos3R (Ours)	✓	–	32.3	31.5	**47.3**	**33.1**	**25.1**	**53.7**	**53.9**	**39.5**	1.4

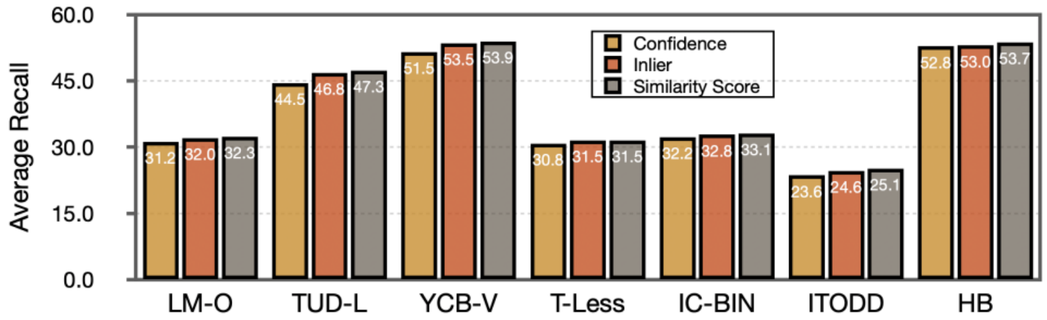
< Comparison with SOTA >

Experiments

- Component Analysis
  - Component ablation
    - Axial rotation 제거 → 전 데이터셋에서 성능 하락
      - ✓Out-of-plane만으론 부족하며, in-plane 보완이 correspondence 품질에 중요
    - MASt3R의 3D-aware 학습이 성능의 핵심 요인
      - ✓MASt3R를 DINOv2, RoMa(dense matcher) 교체 시 큰 폭으로 하락
  - Template selection은 similarity 기준이 가장 안정적
    - AR 기준: Similarity score > # Inlier > Confidence

#	Method	LM-O	TUD-L	YCB-V	T-Less	IC-BIN	ITODD	HB
1	Pos3R	**32.3**	**47.3**	**53.9**	**31.5**	**33.1**	**25.1**	**53.7**
2	*w/o* Ax. Rot.	30.5	43.2	49.3	23.6	21.4	15.4	42.5
3	Face Center	31.5	45.5	53.3	28.0	29.4	22.4	52.5
4	DINOv2 (L11)	18.5	25.0	22.5	14.2	15.6	14.1	32.1
5	DINOv2 (L9)	20.5	26.0	24.0	15.6	16.6	14.8	32.9
6	RoMa [10]	20.8	25.0	32.0	16.2	17.0	15.5	25.1

< Comparison of Pose Estimation Method >



< Comparison of Template Selection Techniques >

감사합니다