A Modern View of 6D Pose Estimation of Novel Objects

2025 하계 세미나

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김수훈

Contents

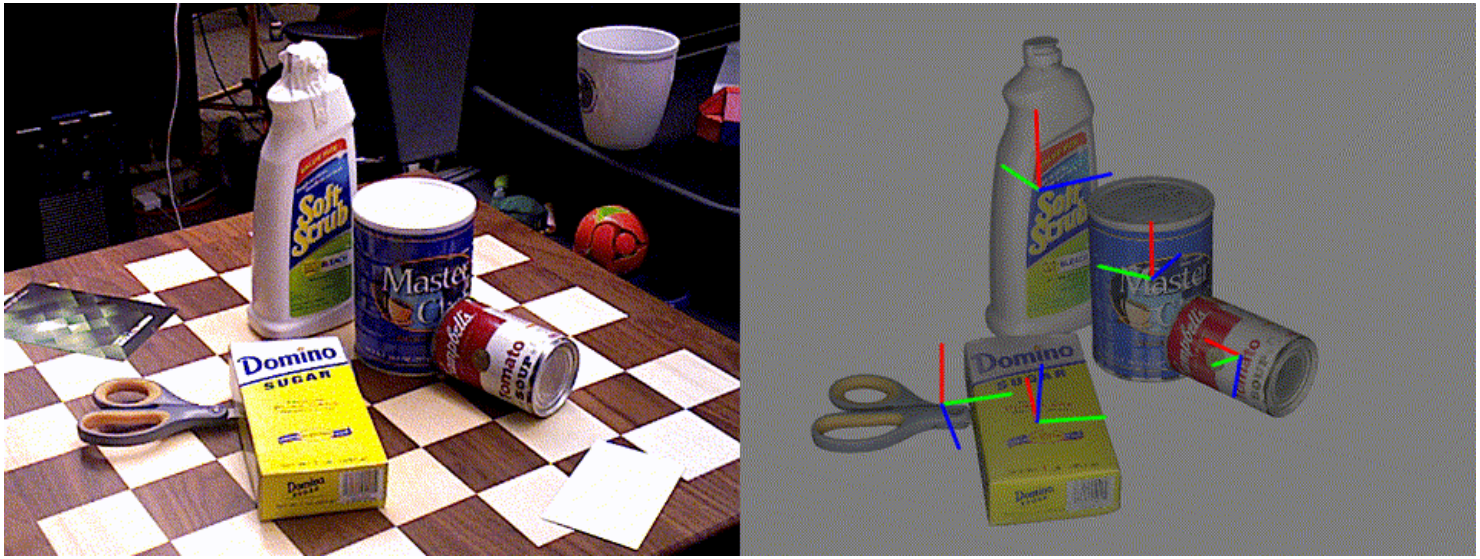
- Introduction

- FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects
  - CVPR 2024
  - NVIDIA

- Any6D: Model-free 6D Pose Estimation of Novel Objects
  - CVPR 2025
  - KAIST & NVIDIA

FoundationPose[1)]

- 6D Pose Estimation
  - 객체의 3D Translation (x, y, z) + 3D Orientation (roll, pitch, yaw) 추정하는 task
  - 2가지 Main setup
    - Model-based
      - 물체의 3차원 형태 정보(CAD 모델)를 확보하고 있다는 전제
    - Model-free
      - 서로 다른 시점에서 촬영된 물체의 몇 장의 참조 이미지가 제공됨



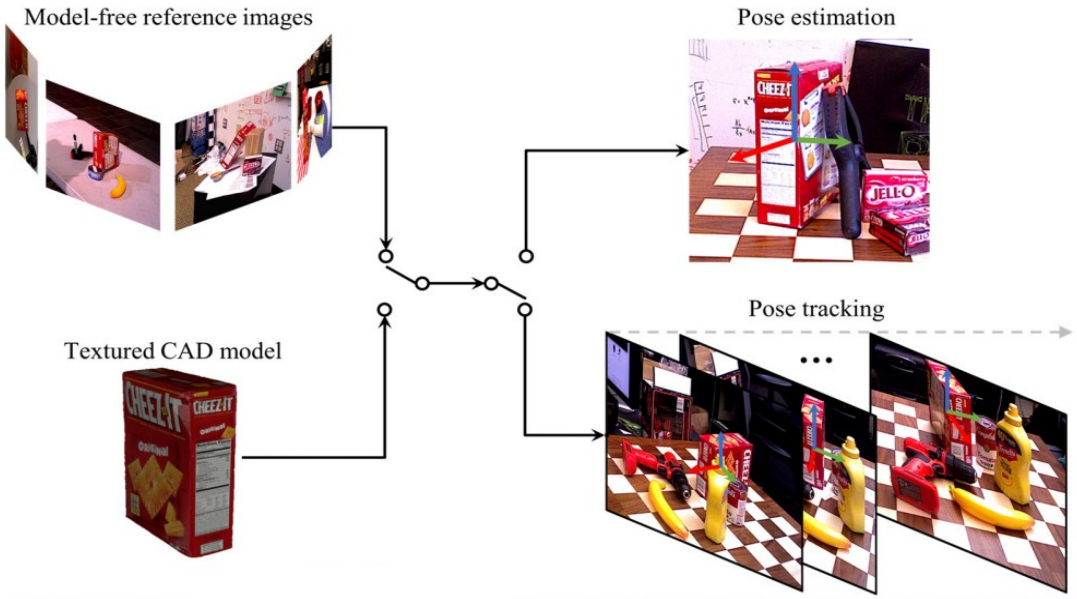
- FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects (CVPR 2024)

Foundationpose[1)]

- Motivation
  - Instance-level
    - 학습 과정에서 경험한 특정 객체 instance 에서만 적용이 가능하며 학습 데이터를 생성하기 위해 texture 포함된 3D CAD 모델을 필요로 함
    - Direct Regression, 2D-3D 대응 관계를 활용한 PnP 알고리즘 등 활용하여 추정함
  - Category-level
    - 특정 instance 에 대해서 학습할 필요는 없으나 여전히 사전 정의된 범주 내의 객체에서 제한적으로 동작함
    - 추가적인 Pose Canonicalization 및 검증 절차가 필요함
      - 3D 물체의 방향을 일정한 기준, reference pose 로 정렬하는 과정
  - Focused on instant pose estimation of arbitrary novel objects
    - Model-based: 객체의 texture 포함한 3D CAD 모델 제공
    - Model-free: 객체의 reference image 제공

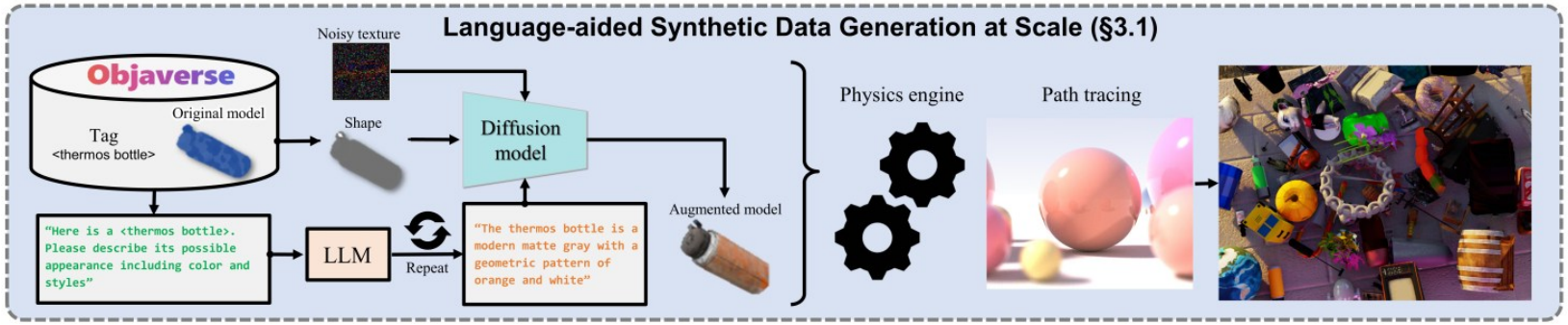
Foundationpose[1)]

- Method
  - A unified framework that performs both pose estimation and tracking for novel objects in both the Model-based and Model-free setups, using RGB-D images
    - LLM-aided synthetic data generation pipeline
    - Novel Transformer-based Architecture
    - Contrastive Learning

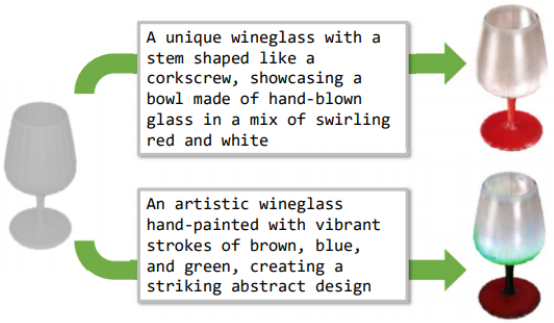


Foundationpose[1)]

- Method
  - LLM-aided Texture Augmentation
    - ChatGPT에 prompt를 제공하여 해당 객체의 가능한 외형을 묘사하도록 요청함
      - 객체와 함께 제공되는 Obj-LVIS tag만 교체
    - ChatGPT의 응답은 Diffusion 모델에 제공되는 text prompt가 되며 자동화하여 대규모의 다양한 데이터셋 생성이 가능함

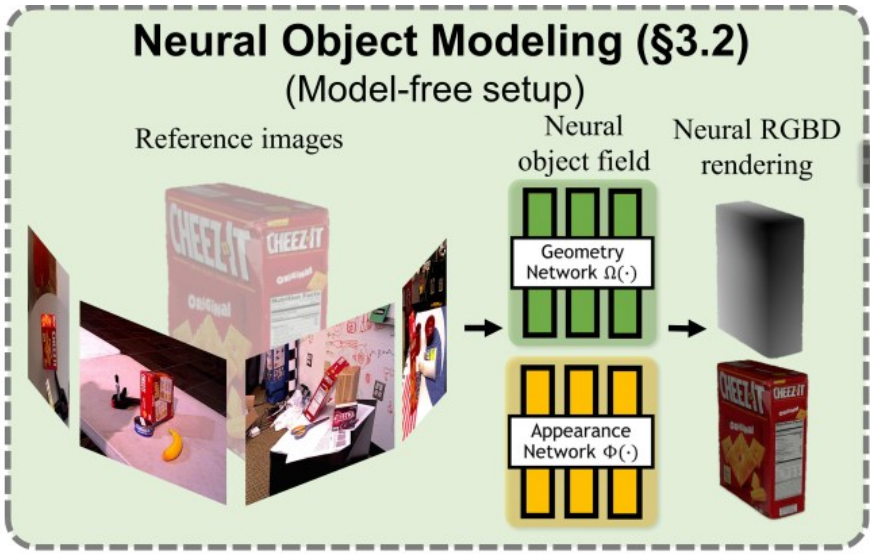


- 동일 객체에 대해서 다른 prompt로 생성된 예시



Foundationpose[1)]

- Method
  - Geometry Network (Ω: x ↦ s)
    - 3D 공간 상의 좌표를 입력으로 활용하여 객체 표면에서 떨어진 거리(s)를 출력함
    - Multi-resolution Hash Encoding 적용하여 좌표를 네트워크에 전달함
  - Appearance Network (Φ: (f_Ω(x),n,d) ↦ c)
    - Intermediate feature vector, 법선 벡터, 시점 방향을 입력으로 활용하여 색상을 도출함



Foundationpose[1)]

- Method
  - Texture Learning
    - 객체 표면 근처의 좁은 영역만을 대상으로 Volume Rendering 진행함

$$c(r) = \int_{z(r)-\lambda}^{z(r)+0.5\lambda} w(x_i)\Phi(f_\Omega(x_i), n(x_i), d(x_i))dt$$

    - SDF = 0 근처일수록 높은 가중치를 부여하여 색상의 품질을 개선함

$$w(x_i) = \frac{1}{1+e^{-\alpha\Omega(x_i)}} * \frac{1}{1+e^{\alpha\Omega(x_i)}}$$

  - Geometry Learning
    - Empty Space loss

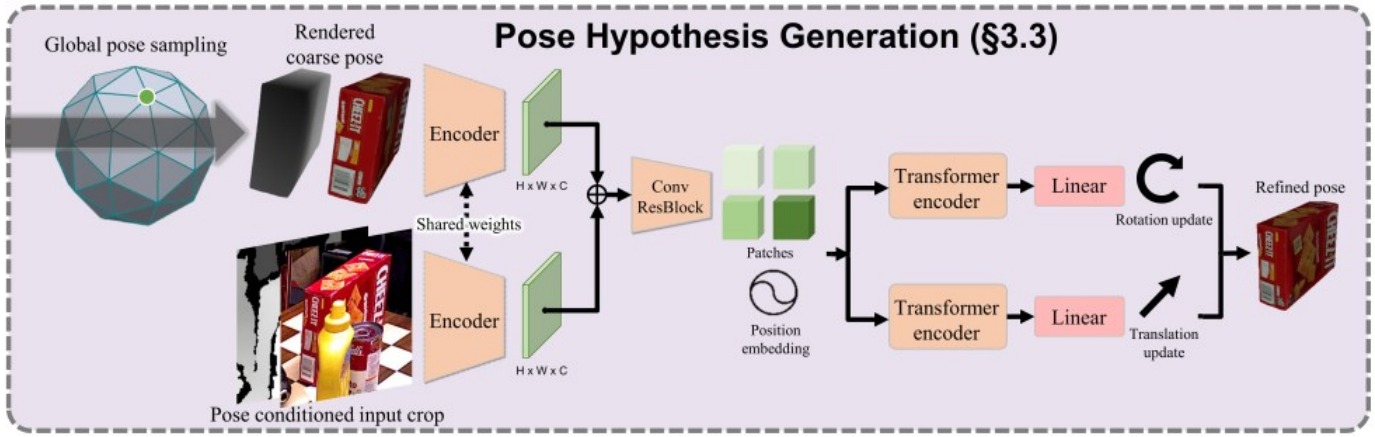
$$L_e = \frac{1}{|R_e|}\sum_{x\in R_e}|\Omega(x)-\lambda|$$

    - Near-surface loss

$$L_s = \frac{1}{|R_s|}\sum_{x\in R_s}(\Omega(x)+d_x-d_D)^2$$

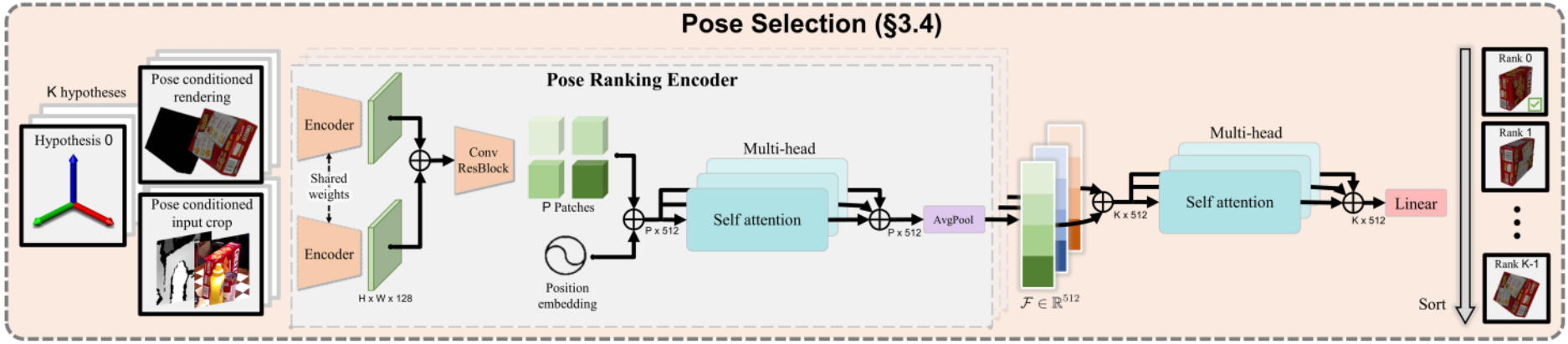
Foundationpose[1)]

- Method
  - Pose Hypothesis Generation
    - 2D-Bbox 내부의 median depth 값을 선택하여 translation 초기화
    - Icosphere 샘플링을 사용한 N_s개 시점에서 N_i개의 in-plane rotation을 통해 rotation 초기화
      - N_s* N_i개의 coarse pose hypotheses 생성
  - Pose Refinement Network
    - Coarse pose 기반으로 객체 중심과 가장 먼 두 점 사이의 거리를 구하여 crop 의 중심과 크기를 결정함
    - 카메라 좌표계에서 Δt, ΔR 반복적으로 계산하여 자세 품질을 개선함



Foundationpose[1)]

- Method
  - Pose Selection
    - Pose Ranking Encoder
      - Cropped 입력 이미지와 렌더링 이미지 간의 비교를 진행하여 두 이미지 간의 alignment 품질을 나타내는 feature embedding, $F \in R^{512}$ 출력함
      - Similarity Scalar 로 투영하는 방식은 절대적인 점수를 출력하도록 강제함
    - Second level of comparison
      - 상대적인 품질 비교가 가능하며 sequence 처리하여 k의 길이가 달라지는 경우에도 일반화 가능함



Foundationpose[1)]

- Method
  - Pose Selection
    - Contrast Validation
      - Propose a pose-conditioned triplet loss to train the pose ranking network

✓기준 sample, anchor 개념을 사용하지 않고 positive, negative 간의 직접적인 비교를 진행함

$$L(i^+, i^-) = \max(S(i^-) - S(i^+) + \alpha, 0)$$

      - Positive / Negative Pose Set

✓GT rotation과의 오차 D가 임계값보다 작으면 positive 로 간주하고, 전체 pose 가설 모두 negative 후보

$$V^+ = \{\, \mathrm{i} : \mathrm{D}(R_i, \bar{R}) < d \,\}, \qquad V^- = \{0, 1, 2, \ldots, K-1\}$$

$$L_{rank} = \sum_{i^-, i^+} L(i^+, i^-) \ (i^+ \in V^+, i^- \in V^-, i^+ \neq i^-)$$

Foundationpose[1)]

- Experiments
  - Datasets
    - LINEMOD, Occluded-LINEMOD, YCB-Video, T-LESS, YCBI-nEOAT
  - Metric
    - Area under the curve (AUC) of ADD and ADD-S
    - Recall of ADD that is less than 0.1 of the object diameter, ADD-0.1d
    - Average recall (AR) of VSD, MSSD and MSPD
      - Introduced in BOP challenge

Foundationpose[1)]

- Model-free Pose Estimation Results
  - All methods are given the perturbed GT bounding box as 2D detection for fair comparison

	PREDATOR [27]		LoFTR [52]		FS6D-DPM [21]		Ours	
Ref. images	16		16		16		16	
Finetune-free	✓		✓		✗		✓	
Metrics	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD
002_master_chef_can	73.0	17.4	87.2	50.6	92.6	36.8	96.9	91.3
003_cracker_box	41.7	8.3	71.8	25.5	83.9	24.5	97.5	96.2
004_sugar_box	53.7	15.3	63.9	13.4	95.1	43.9	97.5	87.2
005_tomato_soup_can	81.2	44.4	77.1	52.9	93.0	54.2	97.6	93.3
006_mustard_bottle	35.5	5.0	84.5	59.0	97.0	71.1	98.4	97.3
007_tuna_fish_can	78.2	34.2	72.6	55.7	94.5	53.9	97.7	73.7
008_pudding_box	73.5	24.2	86.5	68.1	94.9	79.6	98.5	97.0
009_gelatin_box	81.4	37.5	71.6	45.2	98.3	32.1	98.5	97.3
010_potted_meat_can	62.0	20.9	67.4	45.1	87.6	54.9	96.6	82.3
011_banana	57.7	9.9	24.2	1.6	94.0	69.1	98.1	95.4
019_pitcher_base	83.7	18.1	58.7	22.3	91.1	40.4	97.9	96.6
021_bleach_cleanser	88.3	48.1	36.9	16.7	89.4	44.1	97.4	93.3
024_bowl	73.2	17.4	32.7	1.4	74.7	0.9	94.9	89.7
025_mug	84.8	29.5	47.3	23.6	86.5	39.2	96.2	75.8
035_power_drill	60.6	12.3	18.8	1.3	73.0	19.8	98.0	96.3
036_wood_block	70.5	10.0	49.9	1.4	94.7	27.9	97.4	94.7
037_scissors	75.5	25.0	32.3	14.6	74.2	27.7	97.8	95.5
040_large_marker	81.8	38.9	20.7	8.4	97.4	74.2	98.6	96.5
051_large_clamp	83.0	34.4	24.1	11.2	82.7	34.7	96.9	92.7
052_extra_large_clamp	72.9	24.1	15.0	1.8	65.7	10.1	97.6	94.1
061_foam_brick	79.2	35.5	59.4	31.4	95.7	45.8	98.1	93.4
MEAN	71.0	24.3	52.5	26.2	88.4	42.1	**97.4**	**91.5**

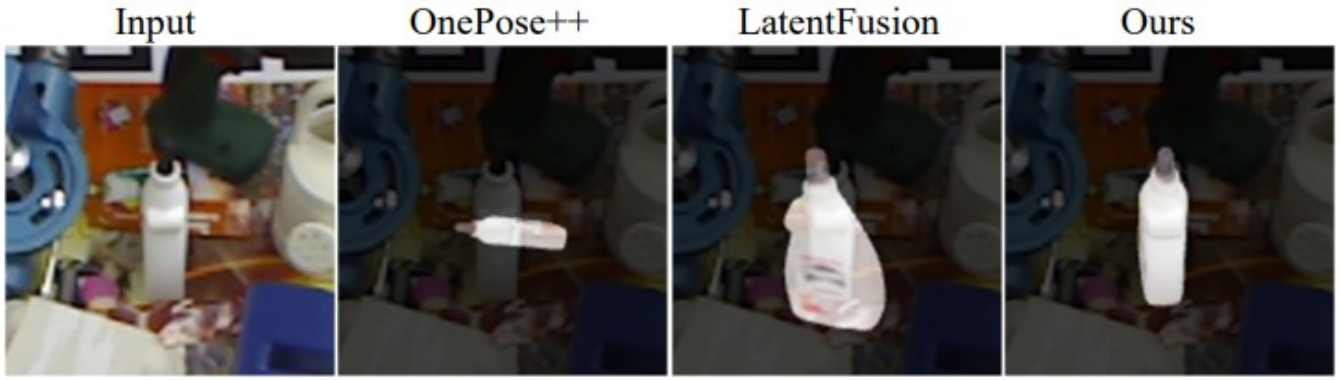
Table 1. Model-free pose estimation results measured by AUC of ADD and ADD-S on YCB-Video dataset. "Finetuned" means the method was fine-tuned with group split of object instances on the testing dataset, as introduced by [21].

Foundationpose[1)]

- Model-free Pose Estimation Results
  - RGB 기반 방법들은 depth 정보가 없기 때문에 더 많은 참조 이미지를 제공받음

Method	Modality	Finetune-free	Ref. images	Objects ape	benchwise	cam	can	cat	driller	duck	eggbox	glue	holepuncher	iron	lamp	phone	Avg.
Gen6D [38]	RGB	✗	200	-	77	66.1	-	60.7	67.4	40.5	95.7	87.2	-	-	-	-	-
Gen6D* [38]	RGB	✓	200	-	62.1	45.6	-	40.9	48.8	16.2	-	-	-	-	-	-	-
OnePose [53]	RGB	✓	200	11.8	92.6	88.1	77.2	47.9	74.5	34.2	71.3	37.5	54.9	89.2	87.6	60.6	63.6
OnePose++ [18]	RGB	✓	200	31.2	97.3	88.0	89.8	70.4	92.5	42.3	99.7	48.0	69.7	97.4	97.8	76.0	76.9
LatentFusion [48]	RGBD	✓	16	88.0	92.4	74.4	88.8	94.5	91.7	68.1	96.3	94.9	82.1	74.6	94.7	91.5	87.1
FS6D [21]	RGBD	✗	16	74.0	86.0	88.5	86.0	98.5	81.0	68.5	**100.0**	99.5	97.0	92.5	85.0	99.0	88.9
FS6D [21] + ICP	RGBD	✗	16	78.0	88.5	91.0	89.5	97.5	92.0	75.5	99.5	99.5	96.0	87.5	97.0	97.5	91.5
Ours	RGBD	✓	16	**99.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.4**	**100.0**	**100.0**	**99.9**	**100.0**	**100.0**	**100.0**	**99.9**

Table 2. Model-free pose estimation results measured by ADD-0.1d on LINEMOD dataset. Gen6D* [38] represents the variation without fine-tuning.



Foundationpose[1)]

- Model-based Pose Estimation Results
  - All methods use Mask R-CNN for 2D detection

Method	Unseen objects	Dataset LM-O	T-LESS	YCB-V	Mean
SurfEmb [15] + ICP	✗	75.8	82.8	80.6	79.7
OSOP [50] + ICP	✓	48.2	-	57.2	-
(PPF, Sift) + Zephyr [45]	✓	59.8	-	51.6	-
MegaPose-RGBD [31]	✓	58.3	54.3	63.3	58.6
OVE6D [2]	✓	49.6	52.3	-	-
GCPose [70]	✓	65.2	67.9	-	-
Ours	✓	**78.8**	**83.0**	**88.0**	**83.3**

Table 3. Model-based pose estimation results measured by AR score on representative BOP datasets. All methods use the RGBD modality.

Foundationpose[1)]

- Ablation study
  - Results are evaluated by AUC of ADD and ADD-S metrics on the YCB-Video dataset

	ADD	ADD-S
Ours (proposed)	91.52	97.40
W/o LLM texture augmentation	90.83	97.38
W/o transformer	90.77	97.33
W/o hierarchical comparison	89.05	96.67
Ours-InfoNCE	89.39	97.29

Table 6. Ablation study of critical design choices.

  - Effects of number of reference images
    - 두 지표 모두 12장 정도에서 성능이 포화
    - 4장의 참조 이미지만 주어져도 16장의 참조 이미지를 사용한 FS6D보다 나은 성능을 기록함

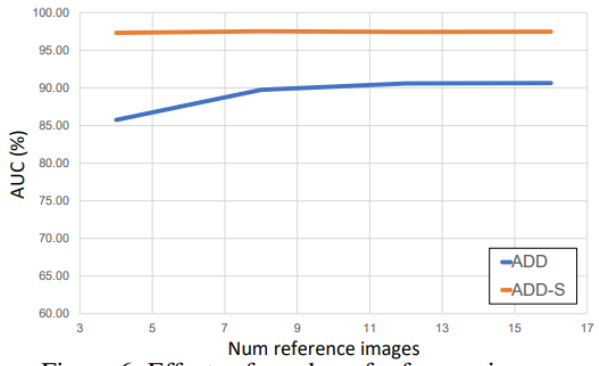


Figure 6. Effects of number of reference images.

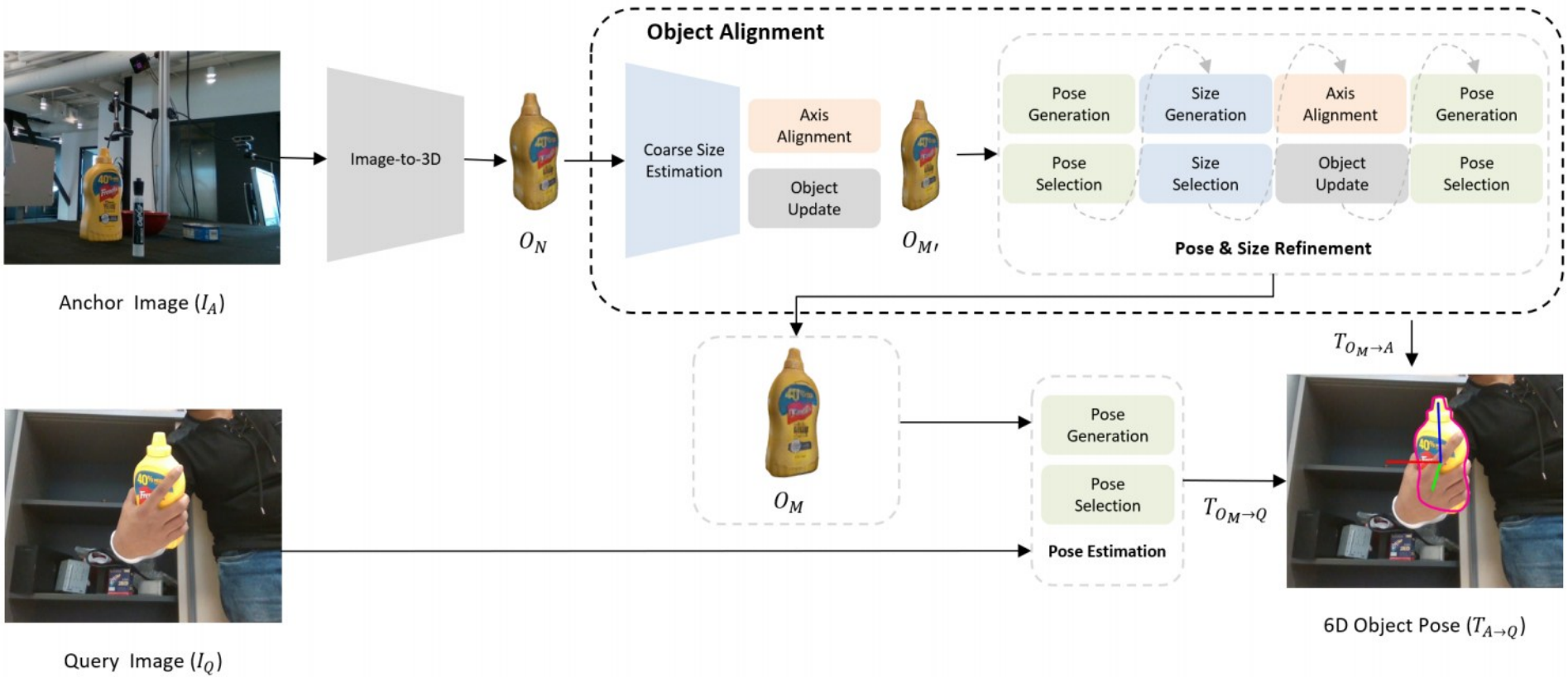
- Any6D: Model-free 6D Pose Estimation of Novel Objects (CVPR 2025)

Any6D[1)]

- Motivation
  - Instance-level
    - 높은 정밀도를 제공하지만 객체의 RGB 텍스처가 있는 CAD 모델에 의존하며 학습 과정에서 경험한 객체에 대해서만 적용이 가능함
  - Category-level
    - Instance-level에서의 한계를 일부 완화하지만 여전히 사전 정의된 객체 class에 제한됨
  - Category-agnostic
    - Model-based
      - Test 과정에서 RGB 텍스처의 3D CAD 모델 필요함
    - Model-free
      - Test 과정에서 다중 뷰의 참조 이미지나 비디오 시퀀스가 필요함
    - Novel Model-free approach
      - Oryon: Language-guidance를 활용하여 single RGB-D 참조 이미지만으로 포즈 추정

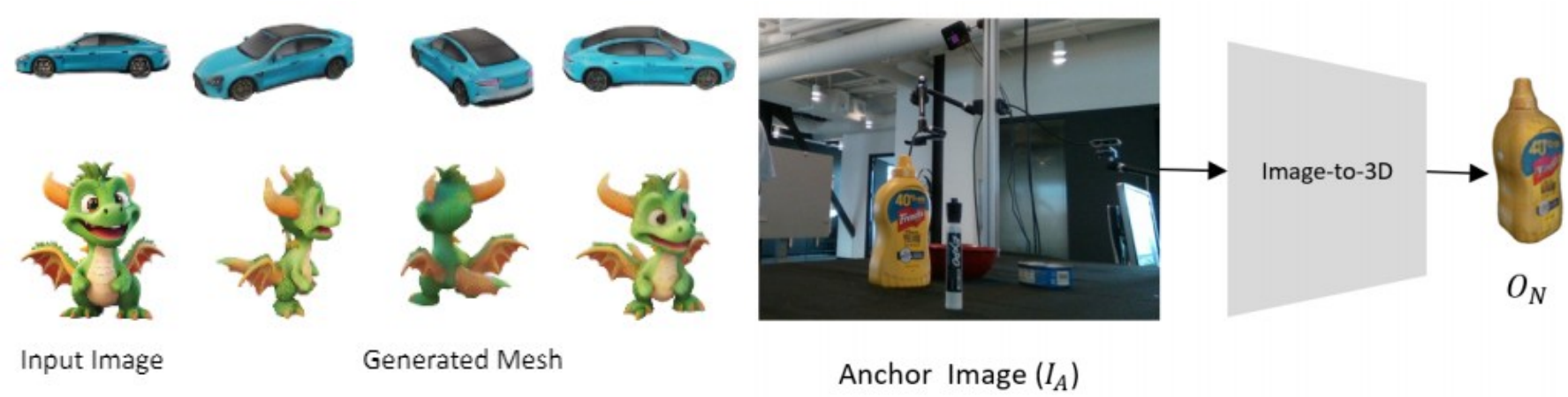
Any6D[1)]

- Method
  - Overview
    - Estimate the relative pose between RGB-D Achor image and an RGB-D Query image



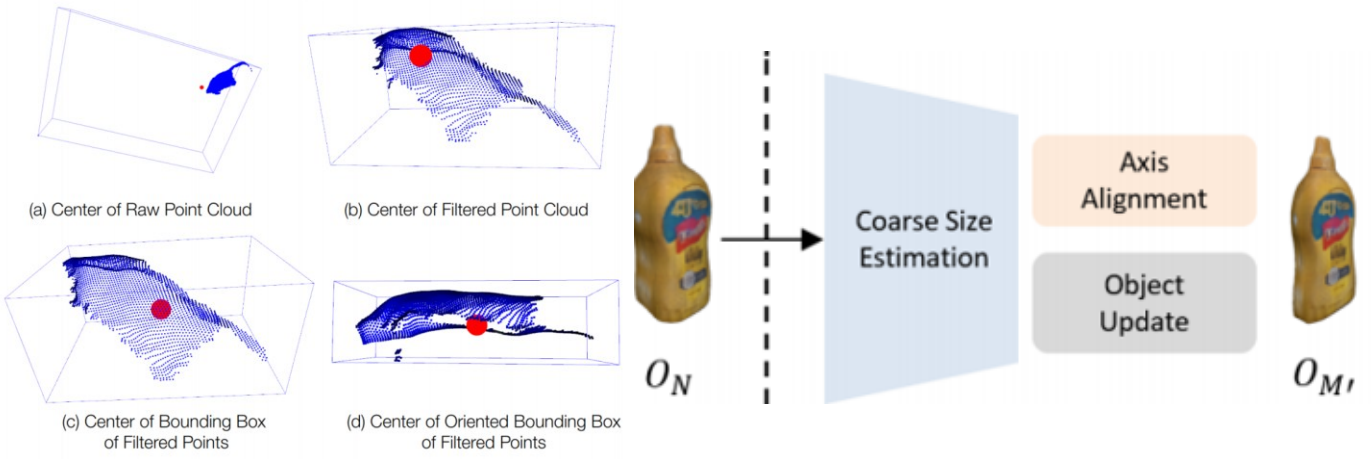
Any6D[1)]

- Method
  - Image-to-3D reconstruction
    - RGB-based Single-view Reconstrcution 에서 좋은 성능을 보인 InstantMesh 채택함
      - Normalized scale 범위 내에서 산출되며 이는 실제 장면에 대해서 정확한 스케일링이나 정렬되지 않았음을 의미함



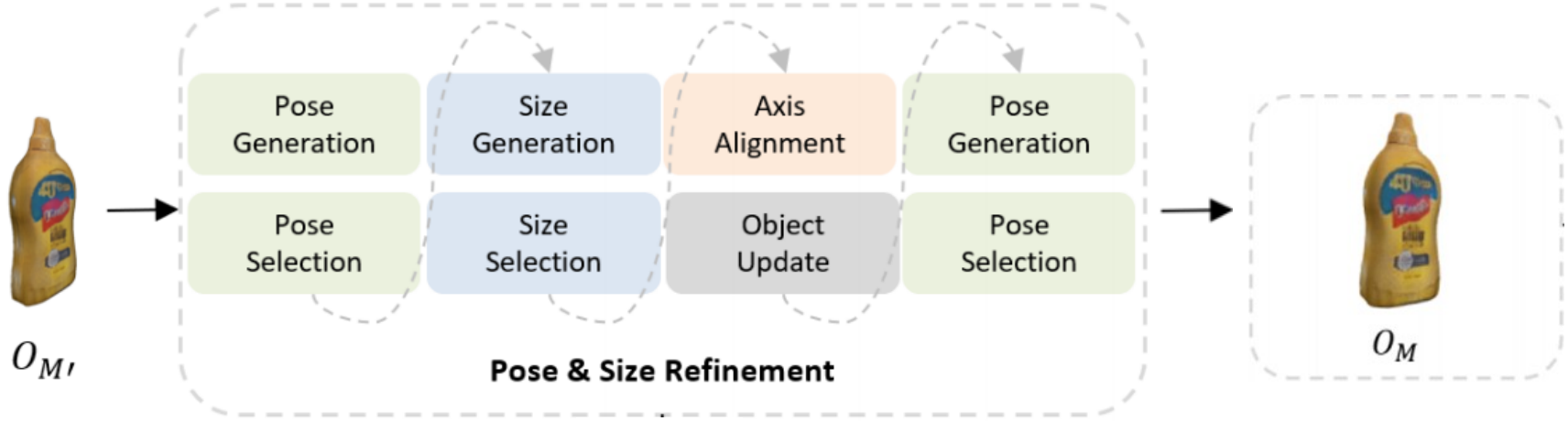
Any6D[1)]

- Method
  - Coarse Object Alignment
    - Estimate object size in a coarse-to-fine manner using Anchor image
      - Comparing point clouds between I_A and O_N from objective center
        - ✓부분적인 시야나 I_A 에 노이즈가 많은 outlier 들이 많은 경우 신뢰하기 어려움
        - ✓Simple axis-aligned bounding box 는 회전을 고려하지 못해 실제 객체의 주축 방향과 misalign 되어 객체 중심을 파악하기 어려움
      - Sample various rotation angles and calculate the IoU
        - ✓Highest IoU is used to transform O_N into a coarsely aligned object shape



Any6D[1)]

- Method
  - Fine Object Alignment
    - Given the coarsely scaled initial shape, jointly refine both the pose and the object size
      - Pose Estimation, Size Estimation, Axis Alignment 모듈로 구성
    - Pose hypothesis generation 단계에서 다양한 크기에 대해서 함께 샘플링을 진행함
      - 각 축에 대해 $\Delta s \in [s_0 = 0.6 , s_1 = 1.4]$ 구간에서 적용

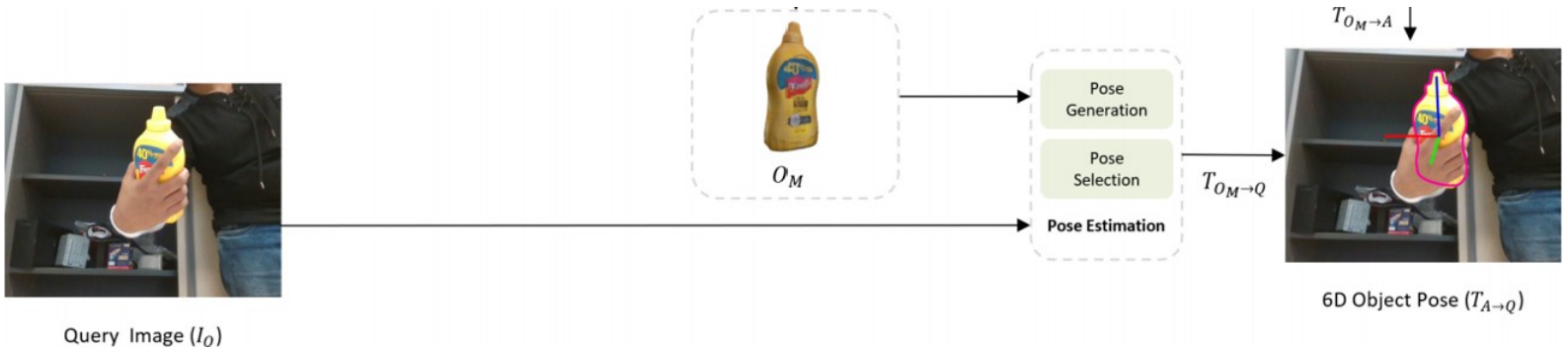


Any6D[1)]

- Method
  - Determine the relative pose $T_A \rightarrow Q$ between an Anchor image and a Query image

$$T_{A \rightarrow Q} = (T_{O_M \rightarrow A})^{-1} \cdot T_{O_M \rightarrow Q}$$

  - Pose Selection
    - Two level render-and-compare strategy
      - Pose Ranking Network evaluates each hypothesis by comparing its rendered view to the cropped observation producing an embedding to quantify alignment quality
      - Apply self-attention to the concatenated embeddings of all hypotheses



Any6D[1)]

- Experiments
  - Datasets
    - Evaluate on five diverse real-world datasets: HO3D, YCBInEOTA, Toyota-Light, REAL275, LINEMOD-Occlusion (LM-O)
  - Metrics
    - Area under the curve (AUC) of ADD and ADD-S
    - Recall of ADD that is less than 0.1 of the object diameter, ADD-0.1d
    - Average recall (AR) of VSD, MSSD and MSPD
      - Introduced in BOP challenge

Any6D[1)]

- Comparison with SOTA
  - 사람 손과의 상호작용 및 occlusion이 포함된 상황에서도 일관된 성능을 보임
    - ADD-S 지표에서 성능이 대부분 100%에 근접한 것은 역동적인 조건에서도 강건한 자세 추정이 가능함을 입증함

Table 1. Model-free pose estimation results measured by AUC of ADD, and ADD-S, AR on HO3D dataset.

	Oryon [11]			LoFTR [59]			Gedi [53]			Ours		
Modality	RGB-D & Language			RGB-D			Depth			RGB-D		
Metrics	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR
AP10	23.8	0.0	0.4	22.5	0.0	1.2	94.4	1.9	3.5	100.0	16.2	22.2
AP11	25.6	0.0	1.3	59.4	15.6	14.8	100.0	55.0	32.3	100.0	73.8	59.0
AP12	21.2	0.0	1.4	12.5	1.2	2.1	99.4	30.6	20.3	100.0	48.8	28.3
AP13	26.2	0.0	0.6	31.9	1.9	1.9	100.0	13.1	8.8	100.0	74.4	45.0
AP14	8.1	0.0	0.0	25.0	0.0	0.0	76.2	0.0	0.6	100.0	35.6	29.7
SM1	24.7	0.0	1.1	52.8	3.4	1.9	82.0	0.0	1.6	86.5	34.8	27.8
SB11	46.1	0.0	2.4	75.4	8.4	15.6	96.4	13.8	12.1	100.0	86.8	68.9
SB13	29.9	0.0	4.2	33.5	0.0	1.9	98.2	11.4	9.4	99.4	64.1	54.6
MPM10	8.3	0.0	0.2	13.4	0.0	0.3	29.9	0.0	3.1	98.7	26.8	31.3
MPM11	33.8	0.0	0.1	26.1	0.0	0.0	35.0	0.0	0.6	100.0	3.2	32.4
MPM12	17.2	0.0	0.1	5.1	0.0	0.0	42.0	0.0	0.4	100.0	1.3	23.5
MPM13	24.2	0.0	0.4	10.2	0.0	0.3	45.2	0.0	1.3	100.0	15.9	30.9
MPM14	9.6	0.0	1.4	15.9	0.0	2.0	35.7	0.0	1.5	98.7	43.9	44.0
MEAN	23.0	0.0	1.0	29.5	2.3	3.2	71.9	9.7	7.4	**98.7**	**40.4**	**38.3**

Any6D[1)]

- Comparison with SOTA
  - Gedi 결과와 비교해 현저히 뛰어난 성능을 보였고 ADD 수치의 큰 향상은 정밀한 자세 추정에 강하다는 것을 의미함

Table 2. Model-free pose estimation results measured by AUC of ADD, ADD-S, and AR on YCBINEOAT dataset.

	Oryon [11]			LoFTR [59]			Gedi [53]			Ours		
Modality	RGB-D & Language			RGB-D			Depth			RGB-D		
Metrics	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR	ADD-S	ADD	AR
sugar_box1	44.0	0.0	1.1	47.3	0.0	0.1	95.6	0.0	1.7	96.7	14.3	11.3
sugar_box_yalehand0	34.7	3.0	5.2	41.6	0.0	0.1	82.2	6.9	21.5	89.1	75.2	44.4
mustard0	48.6	0.0	3.5	47.3	20.3	15.2	100	0.0	19.1	100	23	32.4
mustard_easy_00_02	36.2	0.0	0.3	23.2	0.0	1.9	78.3	0.0	20.2	78.3	53.6	39.2
bleach0	10.4	0.0	1.5	55.2	0.0	0.9	74.6	0.0	7.7	98.5	68.7	56
bleach_hard_00_03_chaitanya	24.4	6.7	6.1	60	15.6	18.7	66.7	62.2	35.5	73.3	51.1	37.7
tomato_soup_can_yalehand0	32.8	0.0	4.5	10.7	0.0	6.8	60.3	0.0	7.8	70.2	0	14.1
cracker_box_reorient	13.2	0.0	0	26.3	0.0	0	97.4	0.0	1.8	100	60.5	44.2
cracker_box_yalehand0	15.0	0.0	1.2	22.6	0.0	0.2	89.5	0.0	10.4	97.7	63.9	58.2
MEAN	28.8	1.1	2.6	37.1	4	4.9	82.7	7.7	14.0	**89.3**	**45.6**	**37.5**

Any6D[1)]

- Comparison with SOTA

Table 3. Model-free pose estimation results measured by AUC of ADD(-S), AR, MSSD, MSPD, and VSD on the Toyota-Light (TOYL) dataset.

Method	ADD(-S)	AR	MSSD	MSPD	VSD
SIFT [42]	14.1	30.3	39.6	44.1	7.3
Obj. Mat. [14]	5.4	9.8	13.0	14.0	2.4
Oryon [11]	22.9	34.1	42.9	45.5	13.9
Ours	**32.2**	**43.3**	**55.8**	**58.4**	**15.8**

Table 5. Model-free pose estimation results measured by AUC of AR, MSSD, MSPD, and VSD on the Linemod Occlusion (LM-O) dataset.

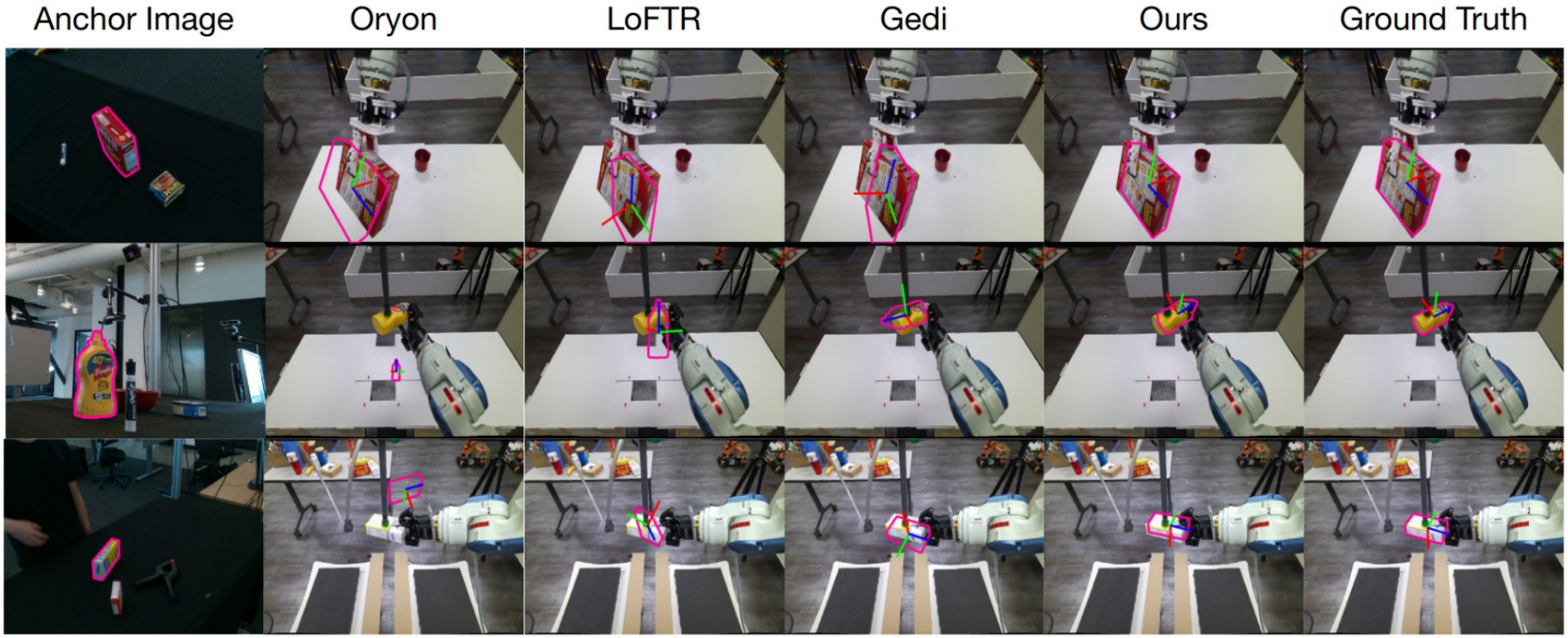
Method	Segmentation	Image-to-3D	Metrics			
			AR	MSPD	MSSD	VSD
GigaPose [47]	CNOS [45]	Wonder3D [41]	17.5	35.8	9.0	7.6
Ours	CNOS [45]	Wonder3D [41]	**28.6**	**36.1**	**32.0**	17.6
Ours	CNOS [45]	InstantMesh [76]	25.2	29.5	27.4	**18.7**

Any6D[1)]

- Qualitative Results

Any6D[1)]

- Qualitative Results



Any6D[1)]

- Ablation Study
  - Object Shape & Object Alignment

Table 6. Ablation Studies of Size Estimation on the HO3D dataset.

Method	Object Alignment			Metrics			
	Coarse Size	Refinement	Axis Align	ADD-S (↑)	ADD (↑)	AR (↑)	CD (↓)
Baseline	✗	✗	✗	28.6	0.00	0.20	1.02
(1)	✗	✗	✗	0.0	0.0	0.0	1.47
(2)	✗	✓	✓	98.0	25.5	26.8	0.53
(3)	✓	✗	✓	83.7	26.6	22.5	0.92
(4)	✓	✓	✗	92.3	23.6	24.9	0.66
Ours	✓	✓	✓	**98.7**	**40.4**	**38.3**	**0.49**

Conclusion

- FoundationPose
  - Unified foundation model for 6D pose estimation and tracking of novel objects, supporting both model-based and model-free setups.
    - Outperforms existing state-of-art methods specially designed for each task

- Any6D
  - Conclusion
    - Novel framework for Model-free object pose estimation that reduces dependence on 3D CAD models and multi-view images
    - Significantly outperforms state-of-the-art methods for occlusions and varying viewpoints
  - Limitation
    - When the initial 3D shape is inaccurate, as our approach does not incorporate shape updating