Category-level 6D Pose Estimation for Rigid and Articulated Objects

2026 동계 세미나

Sogang University

Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By

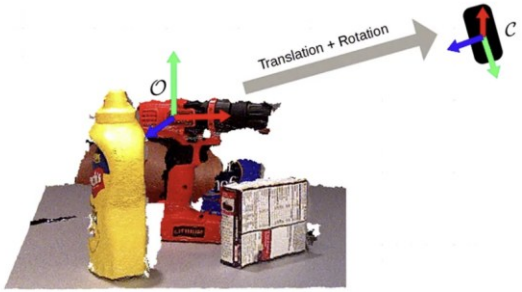
김수훈

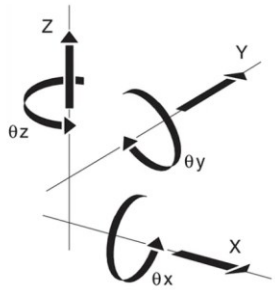
Contents

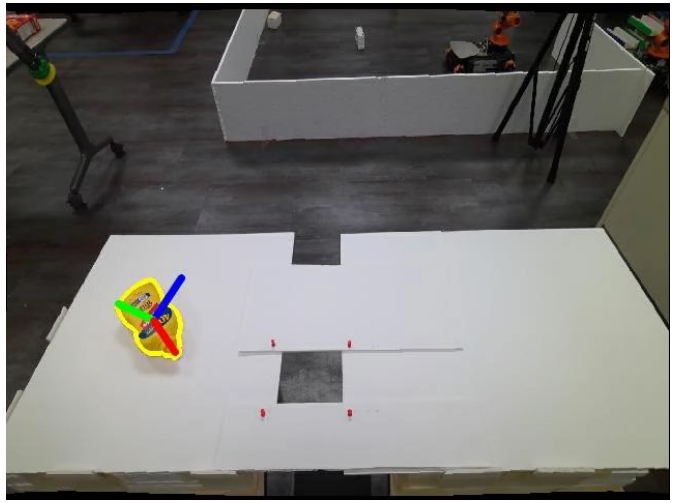
- Introduction
  - 6D Object Pose Estimation
- Preliminary
  - NOCS (Normalized Object Coordinate Space)
- Rethinking Correspondence-based Category-Level Object Pose Estimation
- CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image
- Conclusion

Introduction

- 6D Object Pose Estimation
  - 6 Degrees of Freedom (DoF) = 3 DoF translation (x, y, z) + 3 DoF rotation (orientation)
  - 객체 좌표계에서 카메라 좌표계로의 rigid transformation 계산
  - 로봇의 파지 및 조작을 위한 핵심 인식 모듈로서, 목표 객체의 정확한 자세 추정

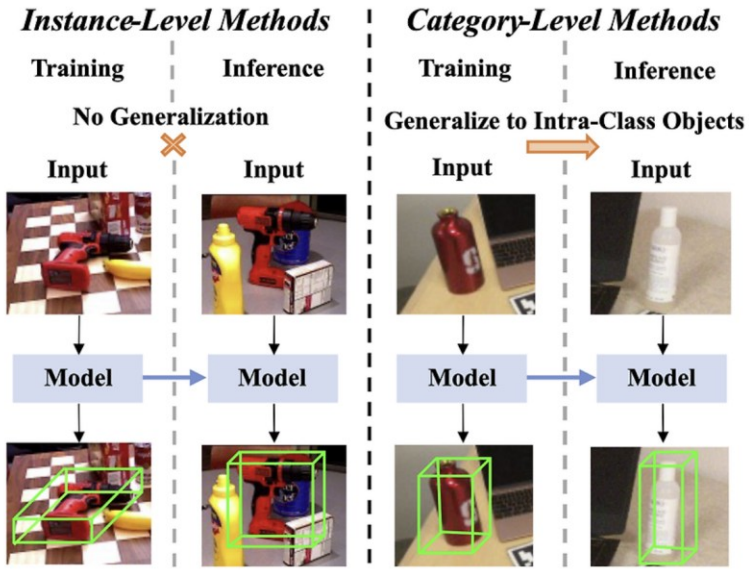






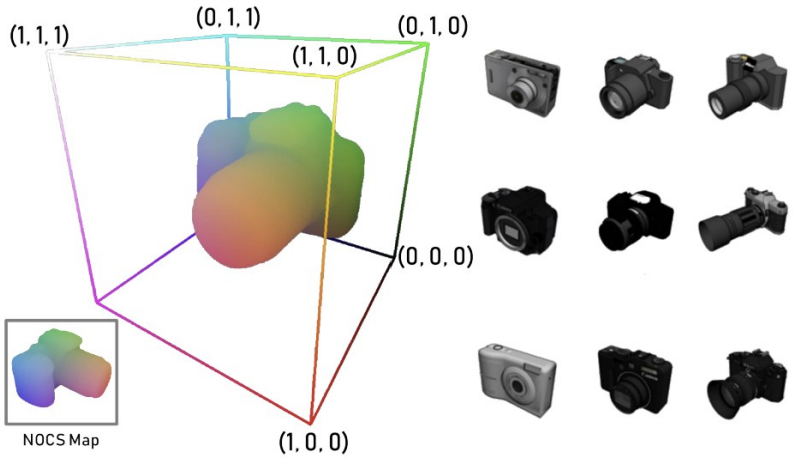
Introduction

- 6D Object Pose Estimation
  - Main setting
    - Instance-level Method
      - 모든 객체의 3D CAD 모델을 사전에 확보한 상태에서 학습 과정에서 본 객체에 대해서만 테스트 한다는 전제
    - Category-level Method
      - 학습 과정에서 본 category 유지하되, 테스트 과정에서는 처음 보는 객체 추정 가능



Preliminary

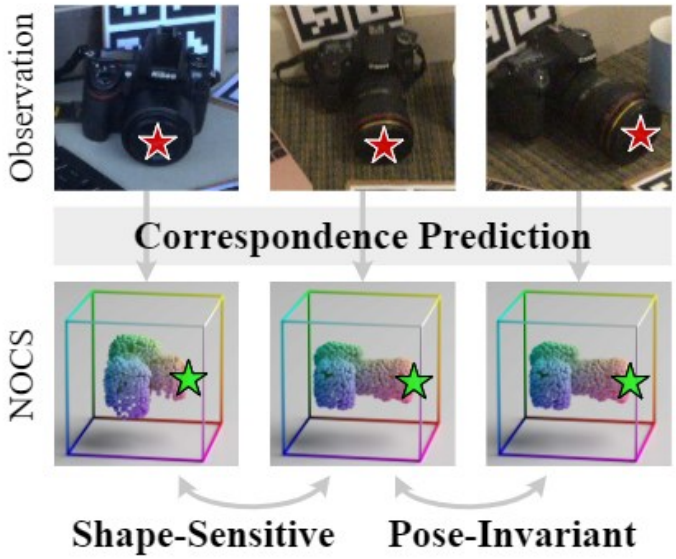
- NOCS (Normalized Object Coordinate Space)
  - 객체를 category별 canonical orientation으로 정렬하고, 3D bounding box 기준으로 크기를 정규화한 canonical 3D 좌표계
    - 같은 카테고리 내에서 의미적으로 대응되는 부분은 유사한 NOCS 좌표를 갖도록 정의
    - 이미지의 각 픽셀에 대해 NOCS 좌표를 예측하고, 실제 관측 3D 점과의 point-wise correspondence로부터 Sim(3) 변환 (R, t, s) 추정
      - 예측된 대응쌍의 outlier를 제거하기 위해 RANSAC을 적용한 뒤, Umeyama 알고리즘으로 두 좌표계 정합



- Rethinking Correspondence-based Category-Level Object Pose Estimation (CVPR 2025)

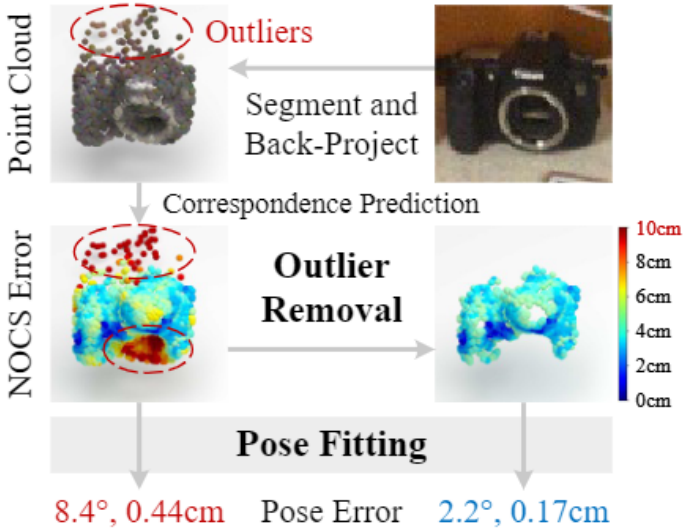
SpotPose[1)]

- Introduction
  - 대부분의 기존 category-level 방법들은 두 단계 correspondence-based 방식 채택
    - 카메라 좌표 공간과 NOCS 사이의 대응 관계 설정
    - Pose fitting 알고리즘을 통해 객체의 포즈 결정
  - Correspondence prediction stage
    - Shape-sensitive and pose-invariant features



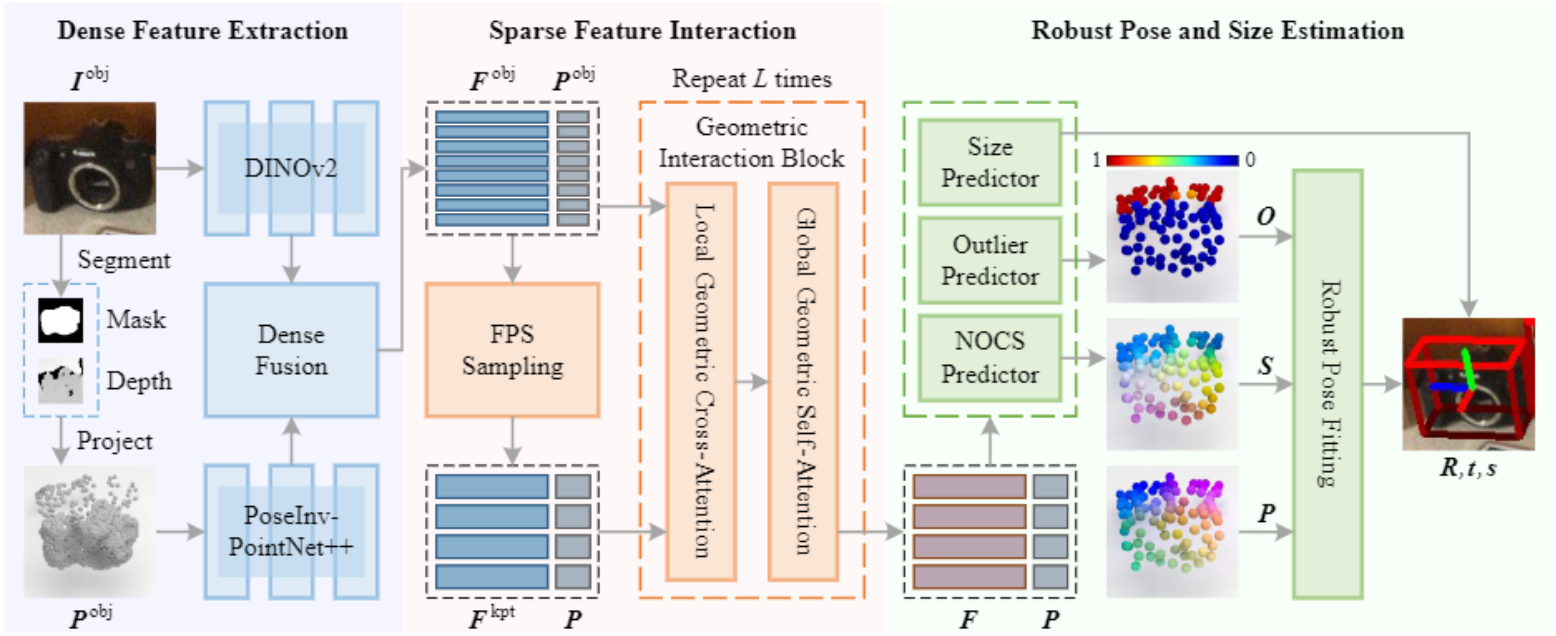
SpotPose[1)]

- Introduction
  - 대부분의 기존 category-level 방법들은 두 단계 correspondence-based 방식 채택
    - 카메라 좌표 공간과 NOCS 사이의 대응 관계 설정
    - Pose fitting 알고리즘을 통해 객체의 포즈 결정
  - Pose fitting stage
    - Removal of outlier correspondences



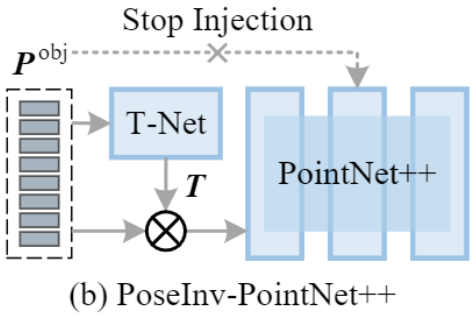
SpotPose[1)]

- Method
  - Cropped RGB-D 이미지, I^{obj} 와 객체 point cloud, P^{obj} 를 입력으로 활용하여 R, t, s 예측

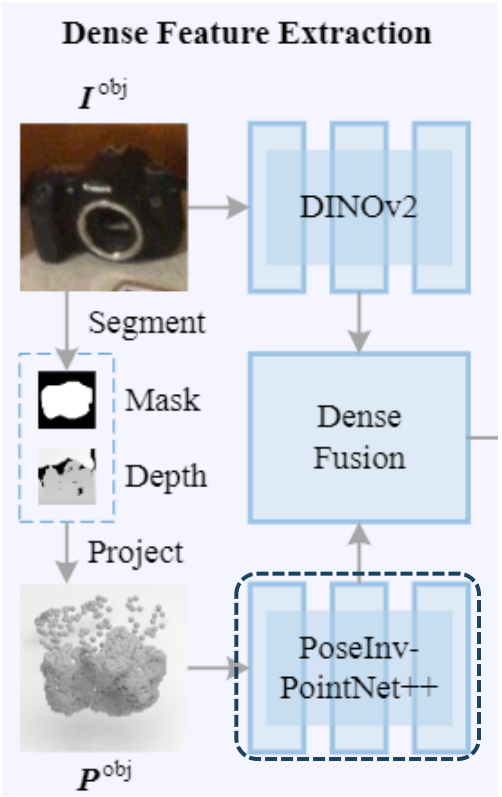


SpotPose[1)]

- Method
  - Dense Feature Extraction
    - Extract pose-invariant geometric features via PoseInv-PointNet++
      - A T-Net is added before PointNet++ to align the input point cloud
      - Injection of absolute coordinates is excluded from the PointNet++
    - For cropped RGB image, employ DINOv2 to extract pose-consistent semantic features

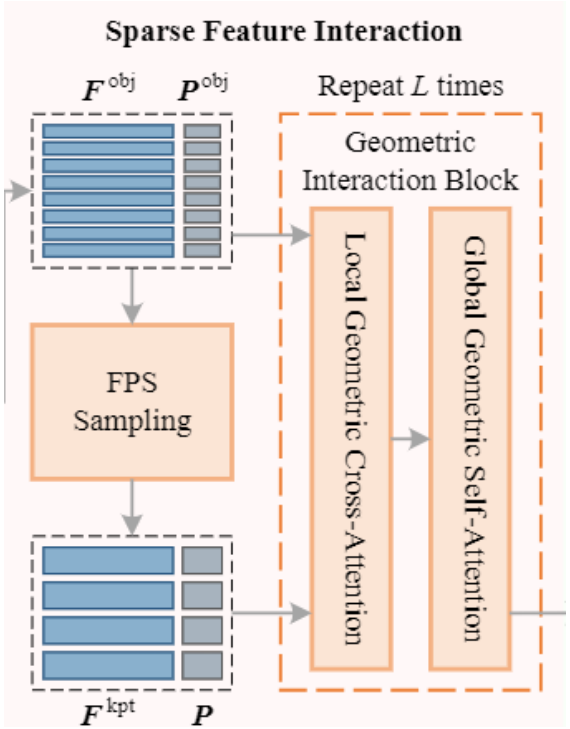


(b) PoseInv-PointNet++



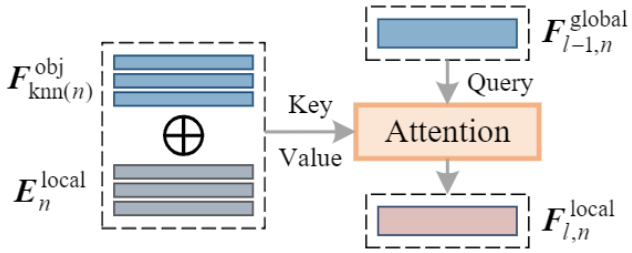
SpotPose[1)]

- Method
  - Sparse Feature Interaction
    - Enhance shape-sensitivity through global feature interaction
    - Represent object shapes with a set of sparse keypoints using Farthest Point Sampling (FPS)
      - Keypoint-wise feature, $F^{kpt} \in R^{N\times 3}$
      - Set of sparse keypoints, $P \in R^{N\times 3}$
    - Distance-and Angle-based Geometric Descriptor
      - Distance embedding, $E^{dis}_{n,m}$
        - ✓$\|P^{att}_m - P_n\|_2 / \sigma_{dis}$
      - K-wise angel embedding, $E^{ang}_{n,m,k}$
        - ✓$\angle(P^{knn}_k - P_n, P^{att}_m - P_n) / \sigma_{ang}$
      - $E_{n,m} = E^{dis}_{n,m} W^{dis} + \max_k \{E^{ang}_{n,m,k} W^{ang}\}$



SpotPose[1)]

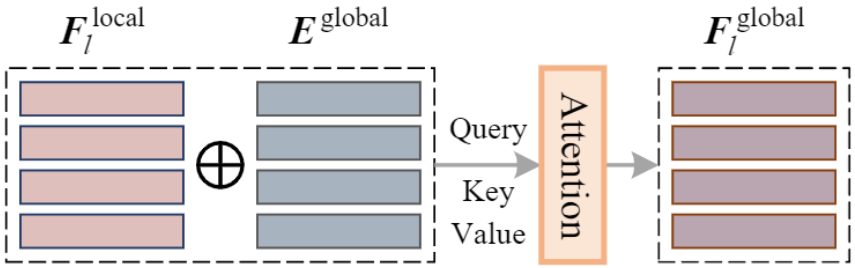
- Method
  - Sparse Feature Interaction
    - Local Geometric Cross-Attention
      - Local context 정보를 keypoint features에 합치기 위한 목적
        - ✓Sparse keypoints, P는 dense points, P^{obj}의 부분집합으로 local geometric 정보를 충분히 표현하지 못함
      - n번째 keypoint, P_n에 대해 P^{obj} 내에서 K_{local}개의 최근접 이웃 선정 후 이에 대응하는 feature, $F_{l,n}^{local}$를 F^{obj}로부터 추출
        - ✓$F_{l,n}^{local} = GCA\left(F_{l-1,n}^{global},\ F_{knn(n)}^{obj},\ E_{l,n}^{local}\right) + F_{l-1,n}^{global}$
        - ✓$GCA(q,C,E) = Attention(q, C+E, C+E)$



(c) Local Geometric Cross-Attention

SpotPose[1)]

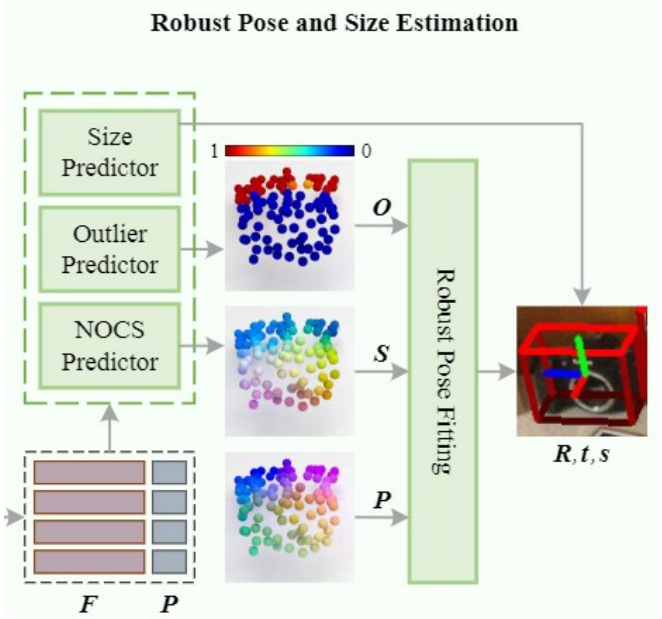
- Method
  - Sparse Feature Interaction
    - Global Geometric Self-Attention
      - Facilitate shape-sensitive global interaction among the keypoint features
      - n번째 keypoint, P_n은 모든 keypoints, $\{P_m|1,\dots,N\}$ 에 대해서 geometric descriptor 계산
        - ✓$F_l^{global} = GSA\left(F_l^{local}, E^{global}\right) + F_l^{local}$
        - ✓$GSA(C,E) = Attention(C+E, C+E, C+E)$
      - Final keypoint feature, $F = F_L^{global} \in R^{N\times D}$



(d) Global Geometric Self-Attention

SpotPose[1)]

- Method
  - Robust Pose and Size Estimation
    - Outlier-aware Correspondence Prediction
      - MLP-based predictor 활용하여 keypoint feature, F로부터 대응되는 NOCS 좌표, S 예측
        - ✓GT NOCS coordinate, $S_n^{gt} = \frac{1}{\|s^{gt}\|_2} (R^{gt})^\top (P_n - t^{gt})$
        - ✓NOCS Loss function, $L_n^{nocs} = \|S_n - S_n^{gt}\|_2$
      - Outlier score prediction
        - ✓Noise in segmentation mask, depth camera sensor
        - ✓Outlier score, $O = Sigmoid\big(MLP([F, MLP(P)])\big)$
    - Pose and Size Estimation
      - Outlier score, O에 0.5 임계값을 적용해서 outlier 제거 후 남은 정상 correspondence만을 사용해서 Umeyama 알고리즘으로 R, t 계산
      - 객체 크기 s는 MLP-based size predictor에 globally averaged keypoint features를 입력으로 활용하여 추정



SpotPose[1)]

- Experiments
  - Datasets
    - REAL275, CAMERA25, HouseCat6D
  - Evaluation Metrics
    - mean Average Precision (mAP) of n°, m cm
    - mAP of 3D Inter-section over Union (IoU_x) at a threshold of x%

SpotPose[1)]

- Experiments
  - Quantitative Results

Method		REAL275						CAMERA25					
		IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
Direct Regression	DualPoseNet [11]	79.8	62.2	29.3	35.9	50.0	66.8	92.4	86.4	64.7	70.7	77.2	84.7
	GPV-Pose [4]	-	64.4	32.0	42.9	-	73.3	93.4	88.3	72.1	79.1	-	89.0
	HS-Pose [37]	82.1	74.7	46.5	55.2	68.6	82.7	93.3	89.4	73.3	80.5	80.4	89.4
	GenPose [36]	-	-	52.1	60.9	72.4	84.0	-	-	79.9	**84.4**	84.6	89.6
	VI-Net [13]	-	-	50.0	57.6	70.8	82.1	-	-	74.1	81.4	79.3	87.3
	SecondPose [3]	-	-	56.2	63.6	74.7	86.0	-	-	-	-	-	-
Correspondence	NOCS [32]	78.0	30.1	7.2	10.0	13.8	25.2	83.9	69.5	32.3	40.9	48.2	64.4
	SPD [27]	77.3	53.2	19.3	21.4	43.2	54.1	93.2	83.1	54.3	59.0	73.3	81.5
	SGPA [1]	80.1	61.9	35.9	39.6	61.3	70.7	93.2	88.1	70.7	74.5	82.7	88.4
	SAR-Net [10]	79.3	62.4	31.6	42.3	50.3	68.3	86.8	79.0	66.7	70.9	75.3	80.3
	DPDN [12]	83.4	76.0	46.0	50.7	70.4	78.4	-	-	-	-	-	-
	IST-Net [16]	82.5	76.6	47.5	53.4	72.1	80.5	93.7	90.8	71.3	79.9	79.4	89.9
	Query6DoF [34]	82.5	76.1	49.0	58.9	68.7	83.0	91.9	88.1	78.0	83.1	83.9	90.0
	AG-Pose [14]	83.7	79.5	54.7	61.7	74.7	83.1	93.8	91.3	77.8	82.8	85.5	91.6
	SpotPose	**84.1**	**81.2**	**59.7**	**64.8**	**81.5**	**88.2**	**94.3**	**92.5**	**80.4**	83.8	**87.7**	**92.2**

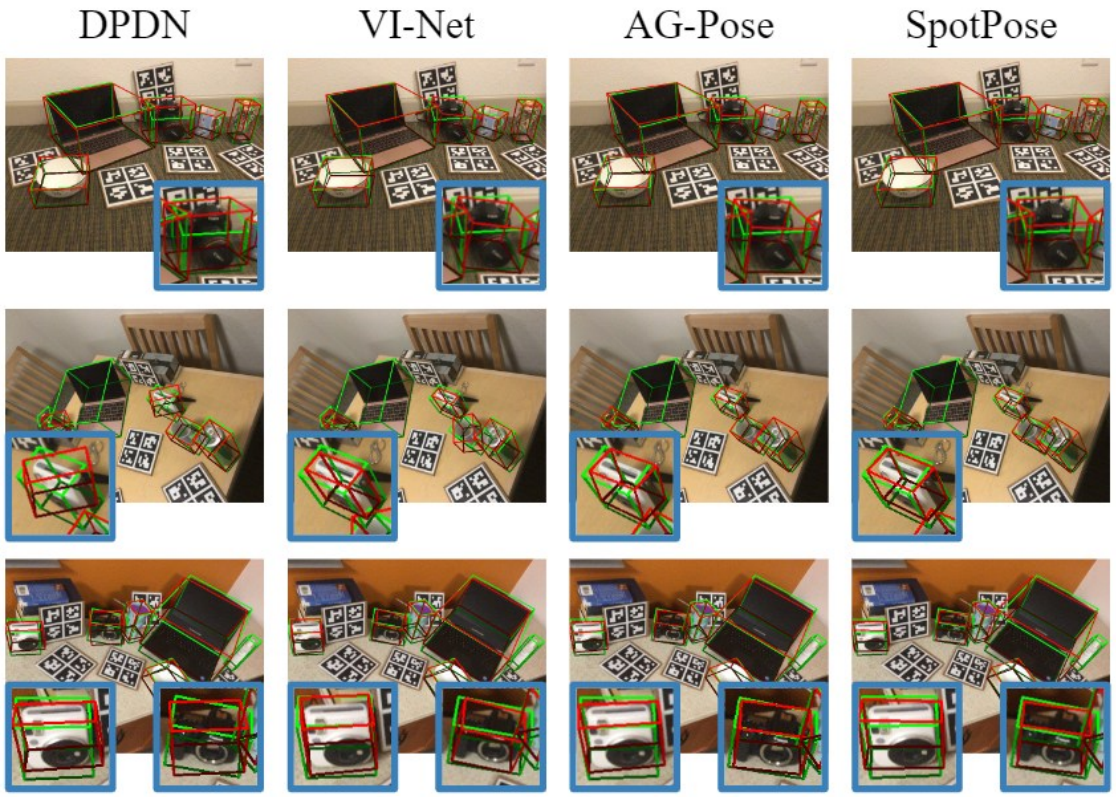
SpotPose[1)]

- Experiments
  - Quantitative Results
    - CAMERA25, REAL275 데이터에 비해 조명·재질·시점·가림·카테고리 다양성이 훨씬 커서 correspondence 기반 pose 추정에 구조적으로 더 어려운 데이터셋

Method	IoU_{25}	IoU_{50}	5°2cm	5°5cm	10°2cm	10°5cm
FS-Net [2]	74.9	48.0	3.3	4.2	17.1	21.6
GPV-Pose [4]	74.9	50.7	3.5	4.6	17.8	22.7
VI-Net [13]	80.7	56.4	8.4	10.3	20.5	29.1
SecondPose [3]	83.7	66.1	11.0	13.4	25.3	35.7
NOCS [32]	50.0	21.2	-	-	-	-
AG-Pose [14]	81.8	62.5	11.5	12.0	32.7	35.8
SpotPose	**89.1**	**77.0**	**23.8**	**24.5**	**52.3**	**54.8**

SpotPose[1)]

- Experiments
  - Qualitative Results



- CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image (CVPR 2025)

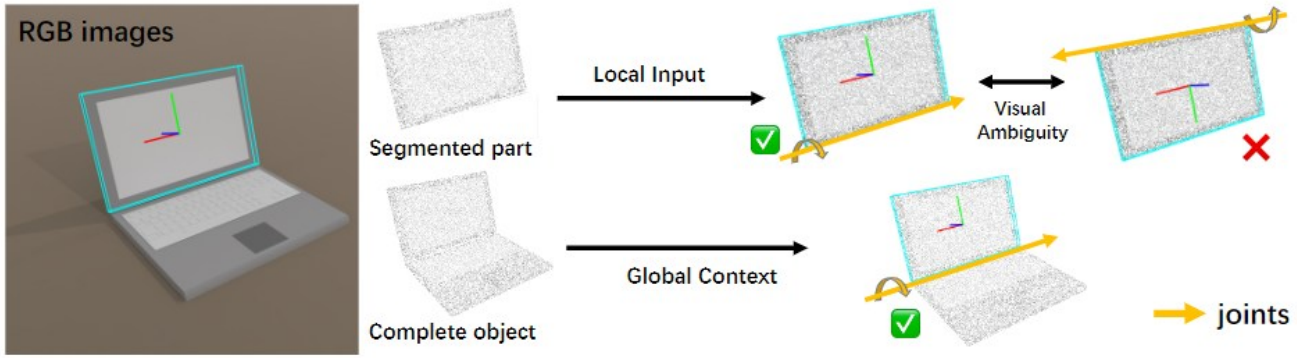
CAP–Net[1)]

- Introduction
  - Rigid 객체에 대한 상태 추정에서는 많은 성과를 내고 있지만, Non-rigid 객체에 대해서는 복잡한 특성으로 인해 여전히 많은 한계가 존재
  - Articulated parts의 경우 여러 관절 구조를 가지기 때문에 인식 오류가 발생할 경우, 관절이 손상될 위험이 있음
    - 로봇 조작 성능을 향상시키기 위해 articulated parts의 인식 및 추정 성능 개선 초점
  - Main challenge
    - Intra-category part variations
      - Often lack specific 3D CAD models, requiring robust intra-category generalization
    - Cross-category contextual variations
      - Diverse contextual configurations across object categories
    - Sim-to-Real domain gap
      - Lack photorealistic RGB images and realistic depth data that simulate real sensor capture

1) Huang, J., Yu, H., Yu, K.-T., Navab, N.,"CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image,"in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

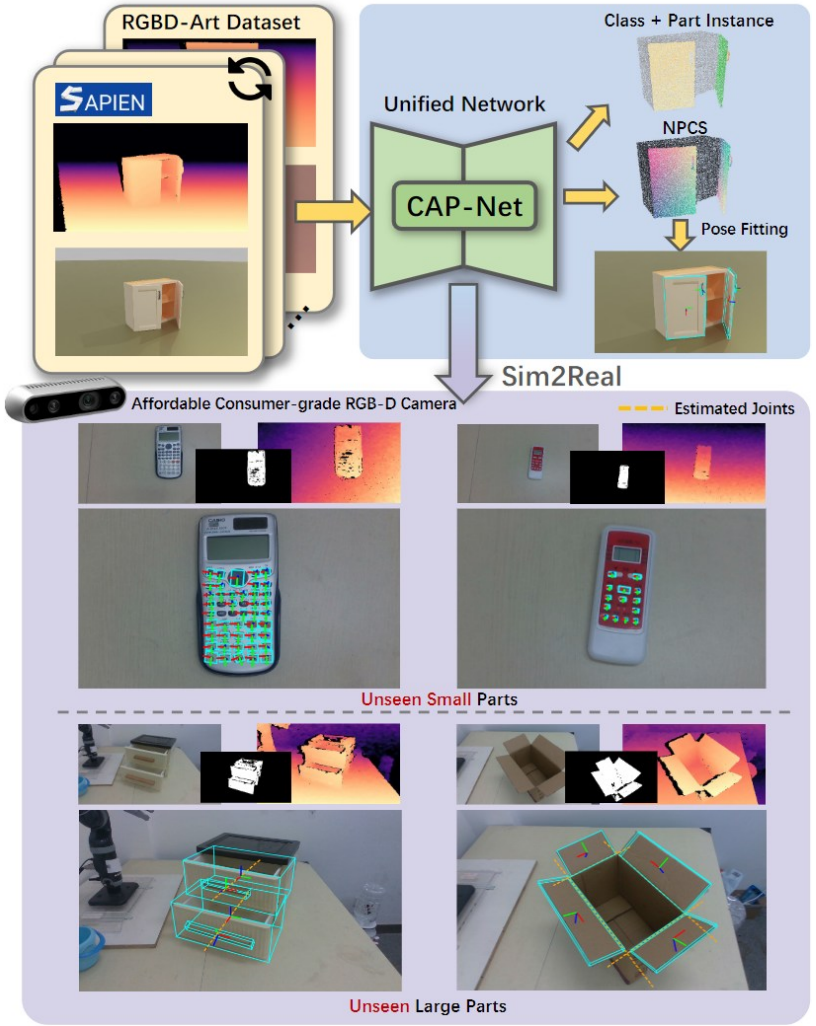
CAP–Net[1)]

- Introduction
  - Point-based method
    - Ignore crucial semantic cues from RGB images, which are vital for precise pose identification
  - Intra-category method
    - Lack cross-category generalization
  - Cross-category method (GAPartNet)
    - 부품을 분할한 후 개별적으로 Normalized Part Coordinate Space (NPCS)를 추정하는 2 stage 방식 사용
    - Accumulate segmentation errors and lose contextual information



CAP–Net[1)]

- Overview
  - Intra-category part variations
    - Utilize NPCS for each part category
  - Cross-category contextual variations
    - Category-agnostic features from a pre-trained visual backbone
  - Sim-to-Real domain gap
    - RGB-D Realistic-Rendering Articulated Object (RGBD-Art) dataset



CAP–Net[1)]

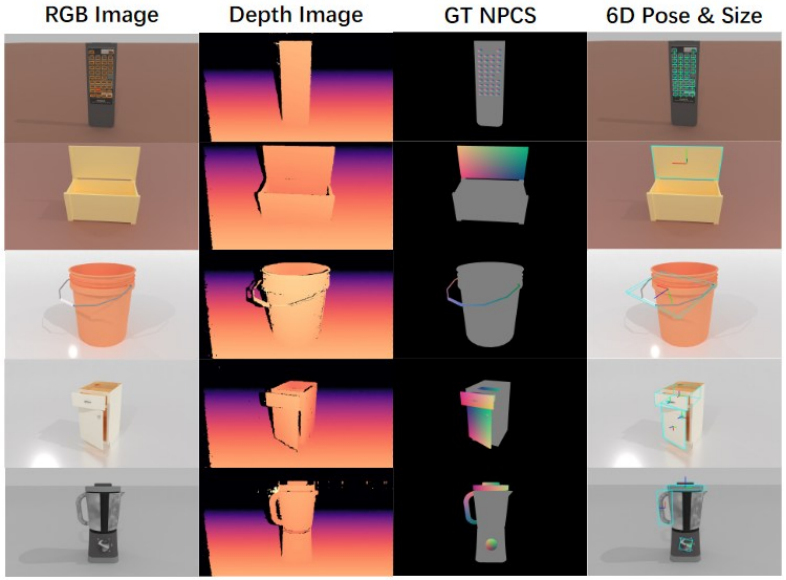
- RGBD-Art Dataset
  - Definition of Poses and Joints
    - Each part category is canonically oriented and normalized to the NPCS, ensuring a consistent definition of pose and joint parameters
  - Limitations
    - Synthetic depth images are idealized
      - Lack realistic noise in real-world sensors
    - Non-photorealistic RGB images

Datasets	Obj.	Img.	Anno.	P-RGB	R-D
ReArtMix [19]	48	100K	-	✔(BG)	✘
GAPartNet [10]	1166	37K	272K	✘	✘
RGBD-Art(Ours)	1045	63K	408K	✔	✔

1) Huang, J., Yu, H., Yu, K.-T., Navab, N.,"CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image,"in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

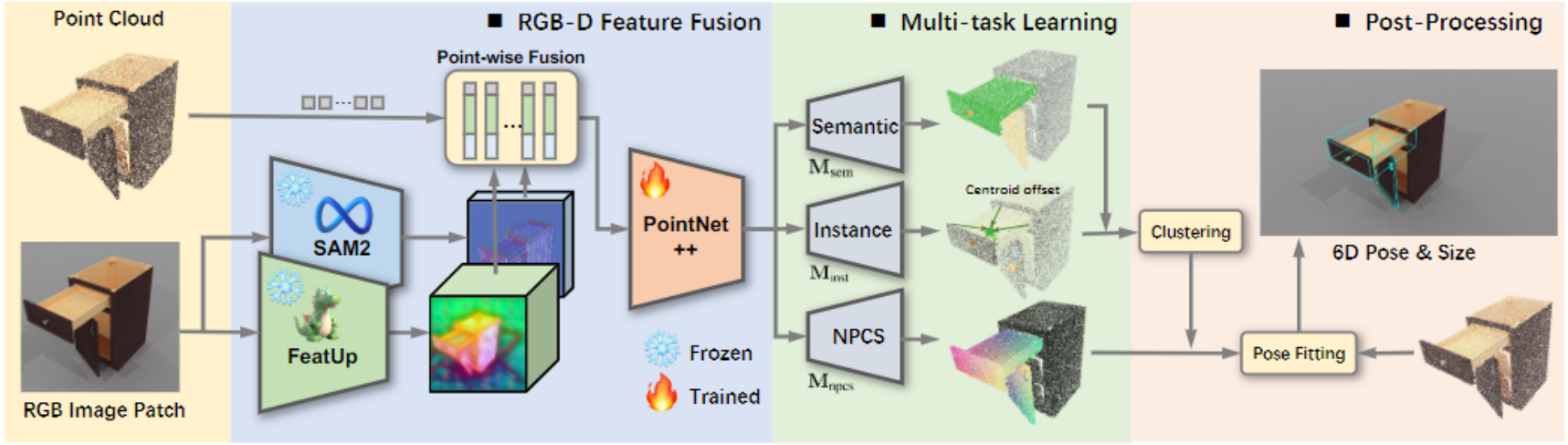
CAP–Net[1)]

- RGBD-Art Dataset
  - Synthesize depth images with realistic sensor noise patterns, simulating an active stereo depth camera similar to the RealSense D415
  - Ray-tracing to achieve photorealism, incorporating domain randomization
  - 9 categories of articulated part types
    - line fixed handle, round fixed handle, hinge handle, hinge lid, slider lid, slider button, slider drawer, hinge door, and hinge knob



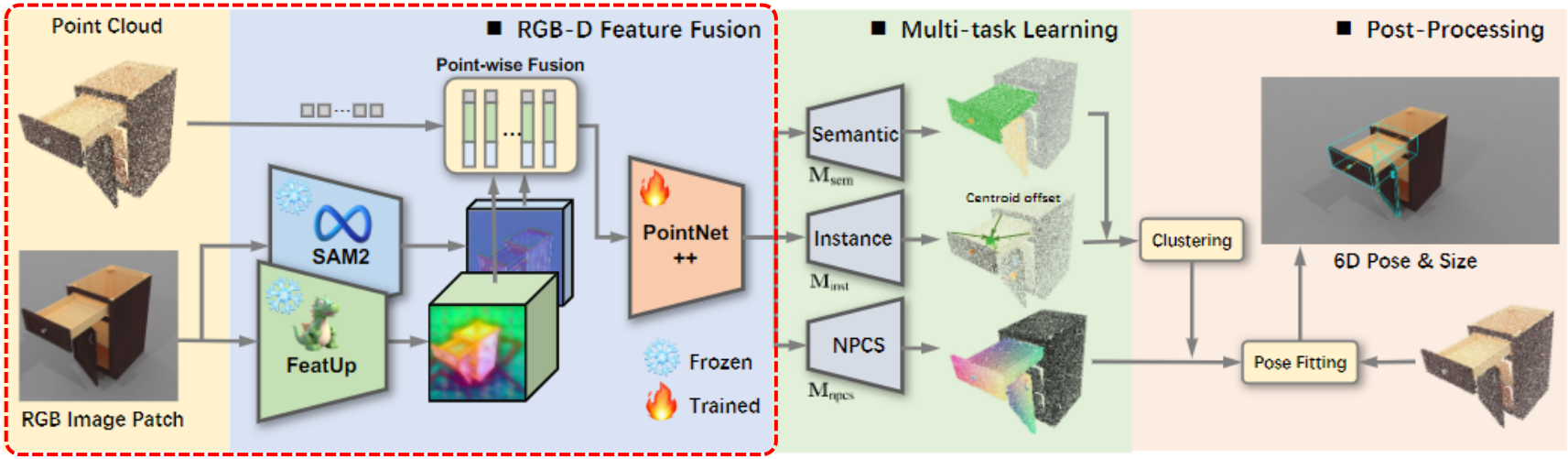
CAP–Net[1)]

- Method
  - Task formulation
    - RGB-D 이미지 patch가 주어졌을 때 point cloud 상의 part semantics를 추정하고 instance 수준에서 구분한 뒤, NPCS 기반 정합을 통해 각 articulated part 6D 포즈와 3D 크기 복원
    - Semantic label을 활용하여 part 클래스 구분 → 동일 semantic label 공유하는 서로 다른 part instance 구분하기 위해 centroid offset clustering → 각 part instance에 대해 추정된 NPCS 기반 정합을 이용해 포즈 및 크기 추정



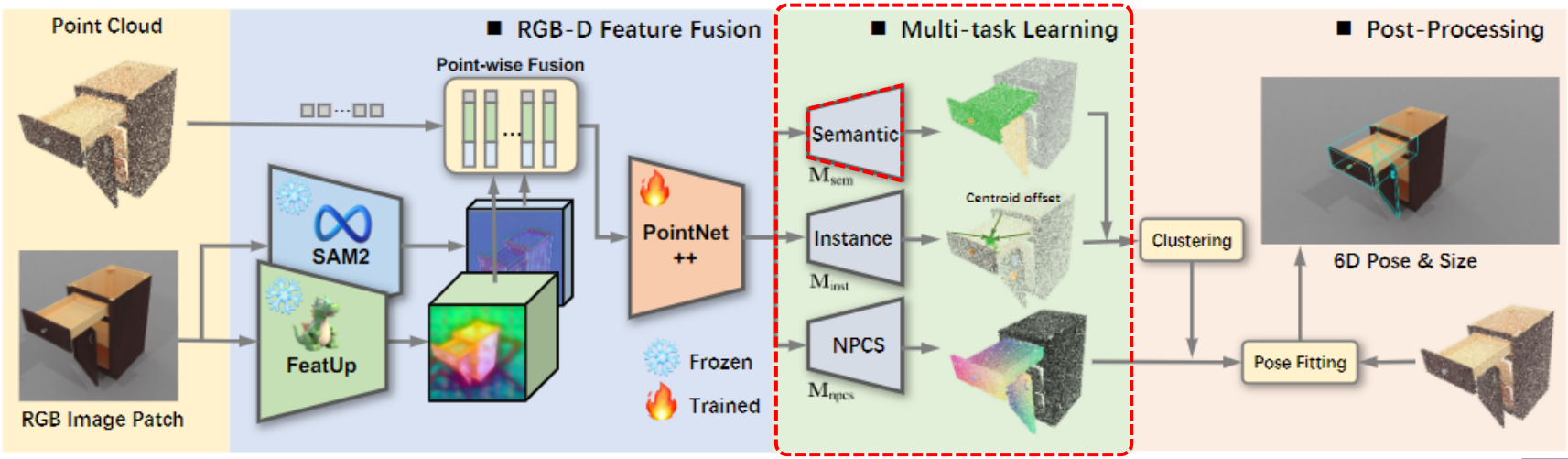
CAP–Net[1)]

- Method
  - Feature Extraction and Fusion
    - Utilizing the pre-trained backbone of SAM2 alongside the FeatUp encoder to extract features from the RGB image patch
      - SAM2: provides strong dense feature (H×W×96)
      - FeatUp: SE(3)-consistent and category-agnostic local semantic feature (H×W×384)
    - Concatenate each RGB feature vector with its corresponding 3D point in a point-wise manner



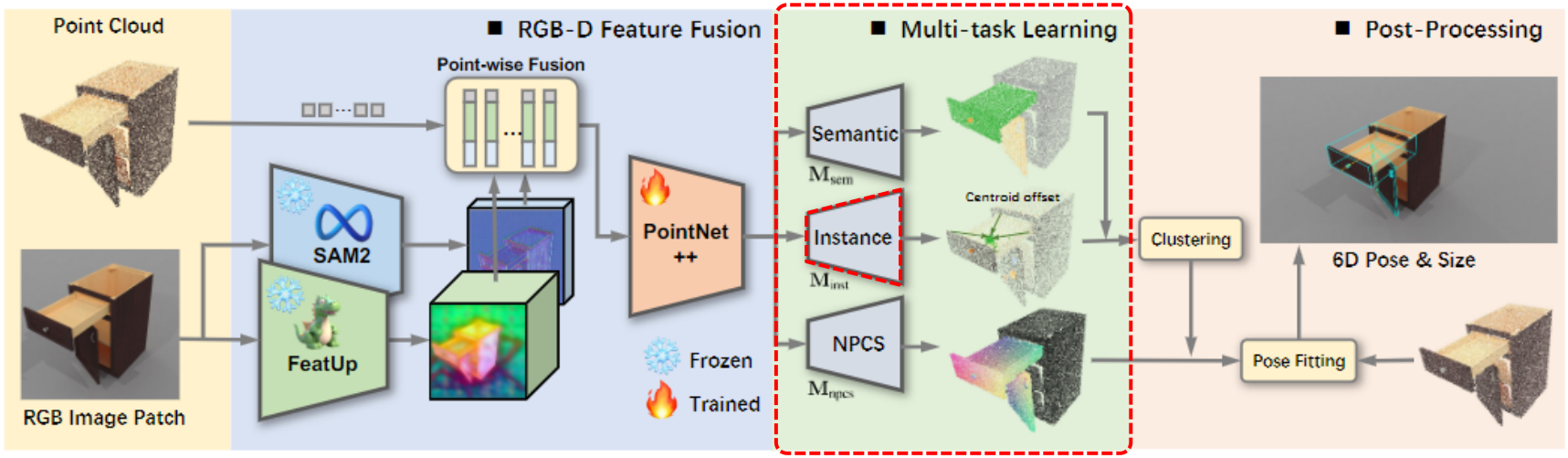
CAP–Net[1)]

- Method
  - Semantic Part Learning
    - GAPartNet는 단일 네트워크를 사용해서 전체 point cloud로부터 가능한 모든 part 분할
      - 포즈 식별에 필요한 global context 손실되며, 각 부품이 이후 독립적으로 처리될 때 segmentation 오류 유발
    - 입력 point (whole object)로부터 semantic label 추정
      - Class와 Instance 분리하여 추정하여 part 구조와 개수가 가변적이여도 일반화 가능
      - 추출된 RGB-D feature 기반으로 각 point의 semantic label 추정



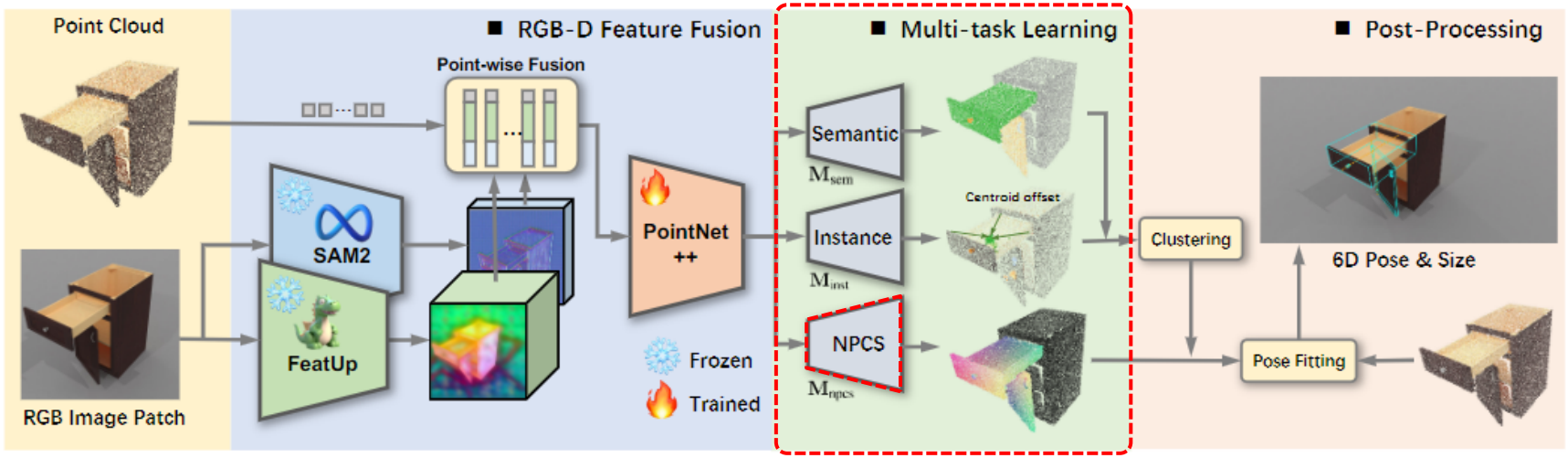
CAP–Net[1)]

- Method
  - Centroid Offset Learning
    - 하나의 객체 내에서 여러 part instance 가 동일한 semantic label을 공유할 수 있는 것을 고려해서 중심을 예측하는 모듈 설계
      - 각 point로부터 해당 part instance 중심까지의 Euclidean translation offset 예측
      - $L_{inst} = \frac{1}{N}\sum_{i=1}^{N}\|\Delta o_i - \Delta\hat{o}_i\| \cdot \mathbb{I}(p_i \in S_k)$



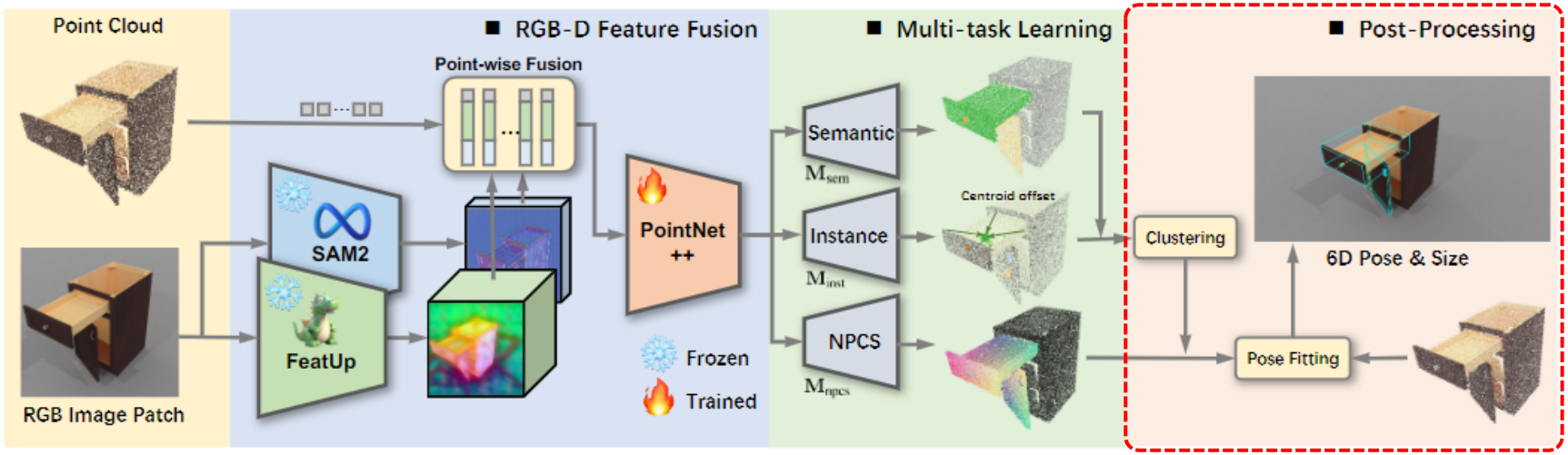
CAP–Net[1)]

- Method
  - NPCS Learning for Pose and Size Estimation
    - 객체 point cloud를 canonical-space point cloud로 mapping 하는 모듈 설계
      - NPCS 좌표를 각 축마다 32개의 bin으로 나눠 classification 문제로 취급하여 해 공간을 줄여 regression보다 효과적
      - $L_{npcs} = \frac{1}{3N}\sum_{i=1}^{N}\sum_{i=1}^{N} SCE(\hat{m}_i^n, m_i^n)$



CAP–Net[1)]

- Method
  - Post-Processing
    - Multi-task loss function
      - $L = \lambda_1 L_{sem} + \lambda_2 L_{inst} + \lambda_3 L_{npcs}$
    - Pose fitting 과정을 통해 최종적으로 객체의 6D pose 및 size 복원
      - RANSAC을 적용하여 outlier 제거한 후, Umeyama 알고리즘을 사용하여 예측된 canonical space point cloud 와 추정된 segmented part point cloud 간 정합



CAP–Net[1)]

- Experiments
  - Evaluation Metrics
    - 3D semantic instance segmentation
      - Average precision at a 50% IoU threshold (AP50)
    - Pose Estimation
      - Rotation error, $R_e(°)$
      - Translation error T_e(cm)
      - Scale error S_e(cm)
      - Translation error along the interaction axis d_e(cm)
      - 3D Intersection over Union (mIoU)
      - Accuracy percentages for thresholds of 5°and 5cm, 10° and 10cm (A_5, A_{10})

CAP–Net[1)]

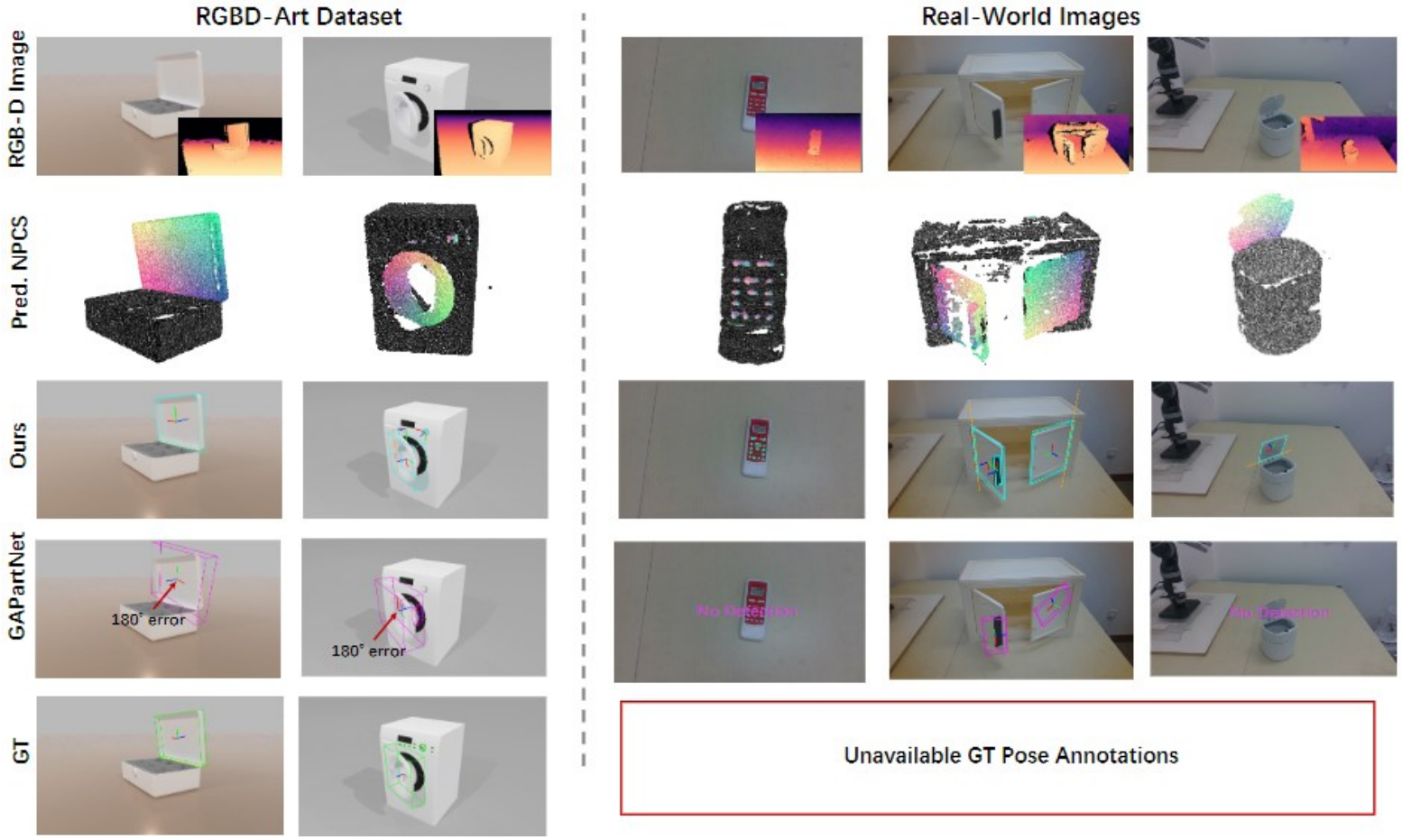
- Experiments
  - Quantitative Results

		Ln.F.Hl.	Rd.F.Hl.	Hg.Hl.	Hg.Ld.	Sd.Ld.	Sd.Bn	Sd.Dw.	Hg.Dr.	Hg.Kb.	Avg.AP50
Seen (%)	SG [32]	0.02	0.07	0.0	10.34	6.24	5.56	14.32	3.51	0.0	4.44
	AGP [20]	1.23	0.15	0.01	13.24	10.21	7.29	16.45	7.41	0.019	6.22
	GAPartNet [10]	3.97	0.26	0.0	25.94	18.41	12.07	26.34	15.15	0.038	11.35
	Ours	**55.30**	**16.88**	**70.71**	**76.05**	**93.76**	**44.06**	**58.09**	**51.05**	**16.23**	**53.58**
Unseen (%)	SG [32]	0.32	5.2	0.0	3.34	0.0	11.5	1.2	6.20	1.12	3.21
	AGP [20]	1.9	0.0	0.0	9.3	0.01	7.8	10.02	6.1	1.5	4.07
	GAPartNet [10]	2.27	0.19	0.0	10.94	0.02	10.42	18.32	14.07	3.1	6.59
	Ours	**28.88**	**0.925**	**0.67**	**51.69**	**1.23**	**28.53**	**20.47**	**24.05**	**18**	**19.38**

Method	$R_e \downarrow$	$T_e \downarrow$	$S_e \downarrow$	mIoU ↑	A_5 ↑	A_{10} ↑
PG [13]	89.30	0.091	0.057	18.41	0.54	1.21
AGP [20]	99.40	0.099	0.061	20.10	0.53	1.32
GAPartNet [10]	83.3	0.061	0.043	39.53	0.71	1.40
Ours	**10.39**	**0.055**	**0.026**	**56.23**	**33.91**	**58.44**

CAP–Net[1)]

- Experiments
  - Qualitative Results



CAP–Net[1)]

- Robotic Experiments
  - Task definition
    - 세 가지 서로 다른 부품 클래스 (drawer, hinge lid, hinge handle)를 선택하고 이에 대응하는 작업은 각각 서랍을 당기는 작업, 뚜껑을 들어 올리는 작업, 그리고 손잡이를 들어 올리는 작업
  - Quantitative Results

	Hinge Handle	Drawer	Hinge Lid	Total
GAPartNet [10]	1/10	2/10	2/10	5/30
Ours	**9/10**	**10/10**	**9/10**	**28/30**

Conclusion

- SpotPose
  - 정확한 pose 추정을 위해 shape-sensitive, pose-invariant feature 추출과 outlier correspondence 제거가 중요

- CAP-Net
  - A unified model for estimating the 6D pose and size of articulated parts at the category-level
  - Pose fitting 과정에서 depth 정보에 의존하기 때문에 표면 point가 누락되는 경우 추정도에 영향을 받을 수 있음