Optical flow-based generalizable 6D pose estimation

2025년 하계 세미나

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

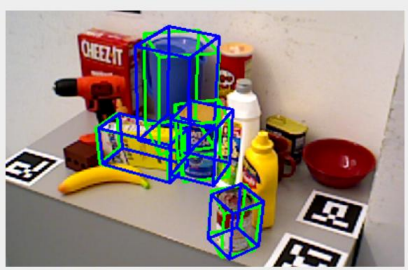
Presented By
장순원

Outline

- Introduction: 6D object pose estimation
- GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects
  - CVPR 2024
  - NAVER AI LABS
- RefPose: Leveraging Reference Geometric Correspondences for Accurate 6D Pose Estimation of Unseen Objects
  - CVPR 2025
  - SNU
- Conclusion

Introduction

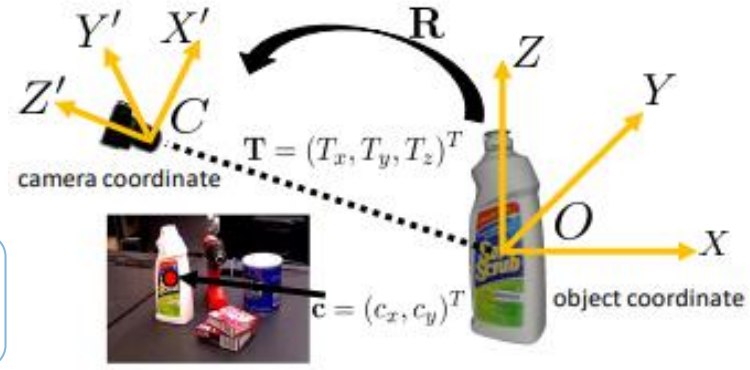
- 6D object pose estimation
  - Camera coordinate와 object coordinate간 translation 및 rotation 추정
  - Challenge: Occlusion, cluttered scene
    - 방향 A에서는 보이던 부분이 방향 B에서는 occlusion 발생으로 관찰이 불가능함



Object detection

$$s \cdot \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = K[R \mid t] \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}$$

6DoF parameter estimation

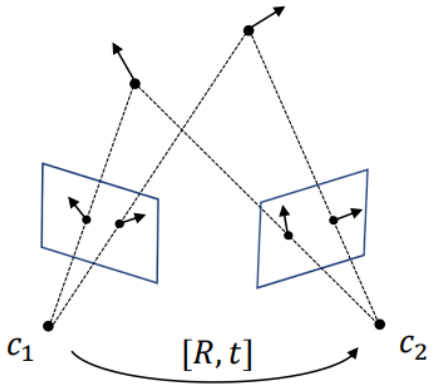


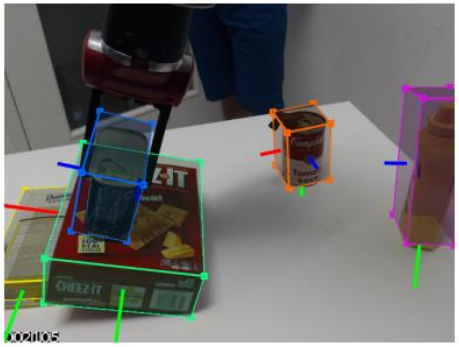
- **3D CAD Model**
- **Camera intrinsic**

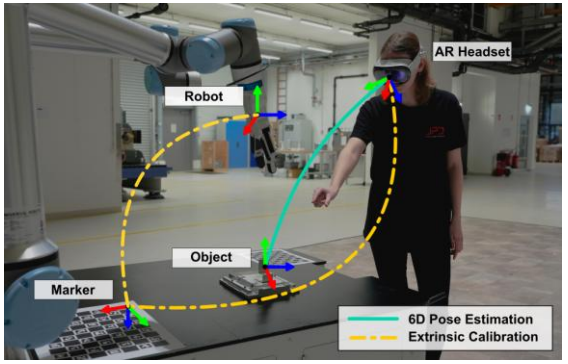
- **Query image**

Introduction

- 6D object pose estimation
  - Application
    - Robot grasping, AR/VR, autonomous driving, etc.
  - Related fields
    - 3D object tracking, Relative pose estimation, SLAM, etc.
  - Current challenges
    - Model-free, Unified model, Zero-shot, Multi-view approaches, etc.

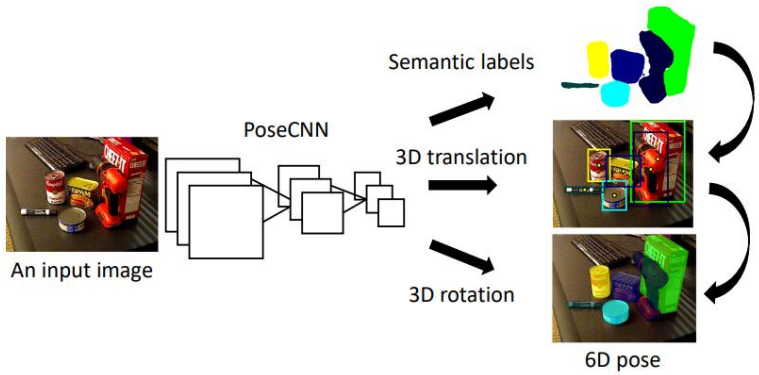




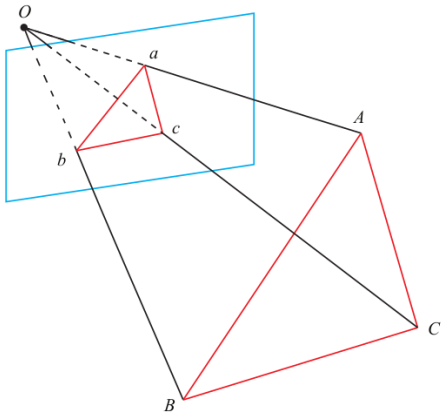


Introduction

- Taxonomy of 6D pose estimation
  - Regression
    - Direct regression of 6DoF parameter with deep learning network
    - Seen data의 시각적 패턴에 의존하여 generalization 성능 저하
    - 새로운 target object의 등장 시 높은 training cost 필요
  - Correspondence-based (key point matching)
    - 2D-3D correspondence key point matching → traditional pose estimation algorithm
  - Render-and-compare



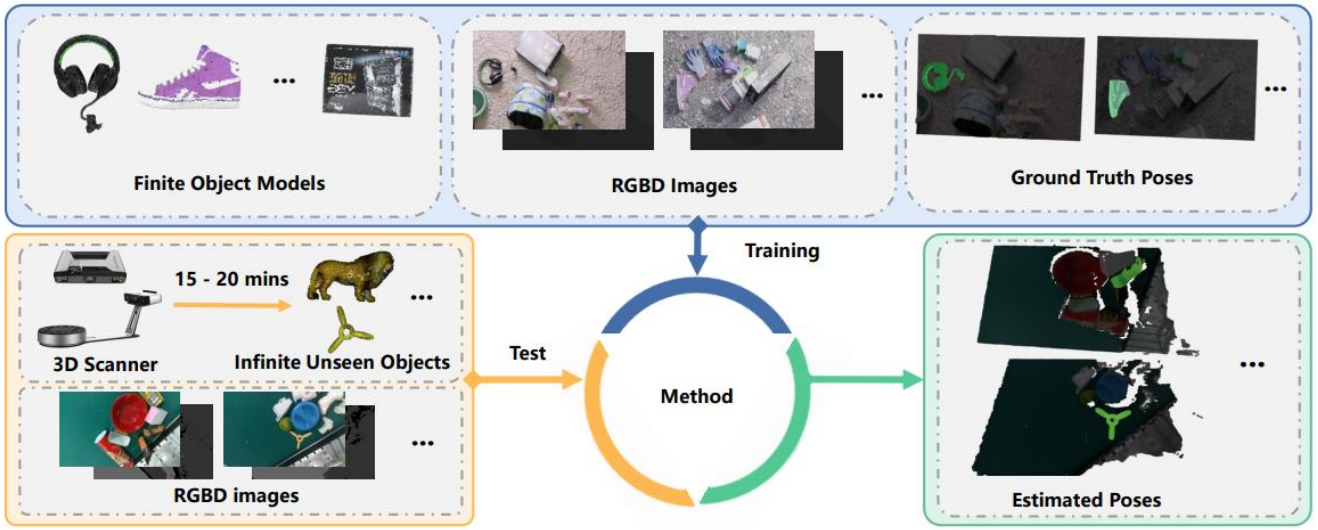
< Regression model – PoseCNN[2] >



< PnP algorithm >

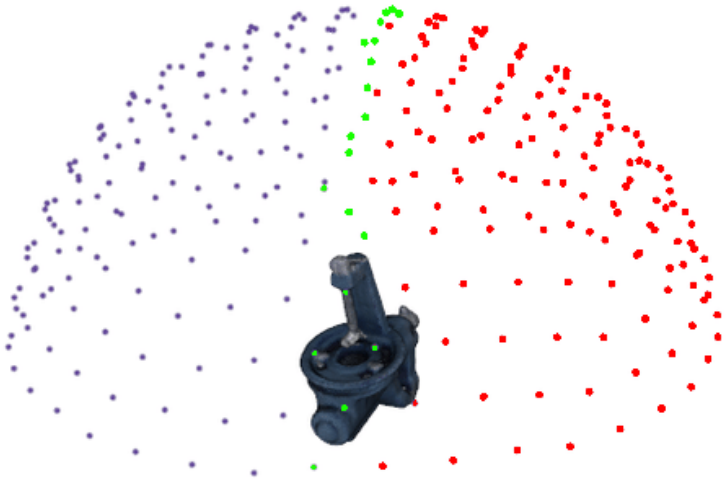
Introduction

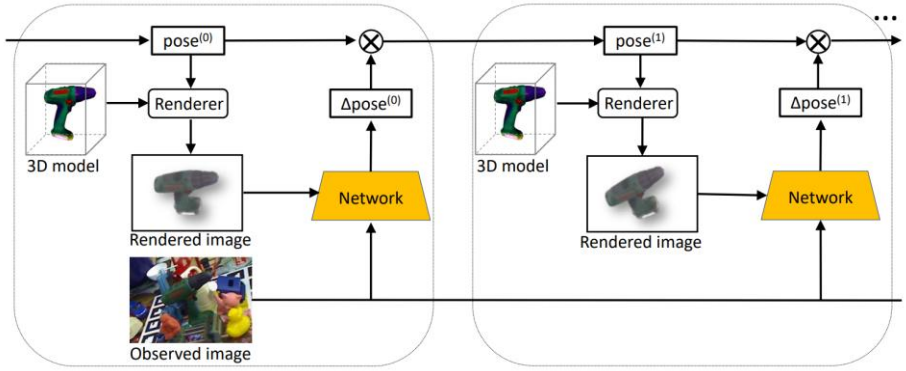
- Generalization to unseen objects
  - 새로운 물체의 3D CAD 모델만 존재한다면 학습 없이 pose estimation이 가능
  - Direct regression과 key point 2D-3D correspondence는 unseen object에서 accuracy 감소
  - 3D CAD model을 reference하여 pose를 추정하는 방법론 필요



Introduction

- Render-and-compare method
  - 현재 추정된 pose로 rendering 된 2D 이미지와 query 이미지를 비교하여 refine
  - Classification → Refinement
    - Pre-rendered template과 비교하여 coarse pose를 classification 형태로 추정
    - Rendered image와 query image를 deep network에 입력하여 pose 변화량 regression
    - Rendered image가 query image와 가까워지도록 반복적으로 업데이트
  - 두 이미지의 관계를 optical flow 형태로 추정하여 활용





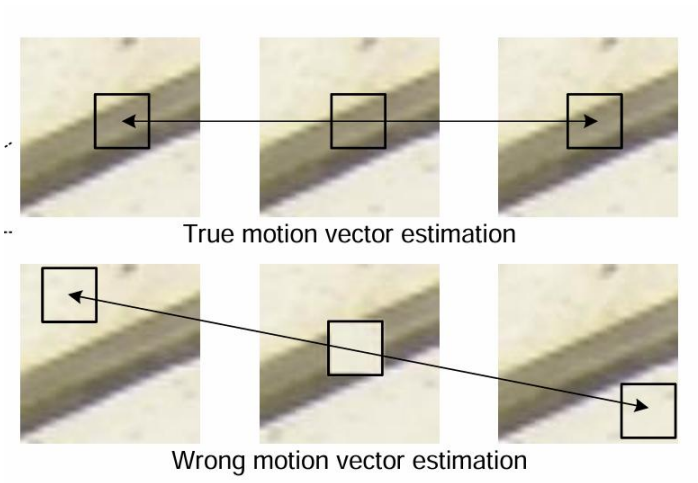
- GenFlow: Generalizable Recurrent Flow for 6D Pose Refinement of Novel Objects (CVPR 2024)

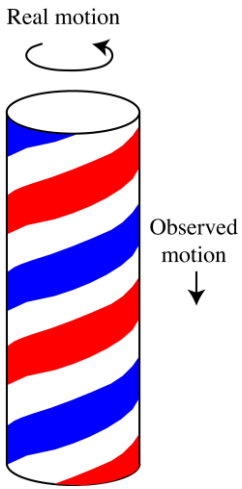
GenFlow[1)]

- 기존의 대표적인 render기반 방식 네트워크
  - OSOP
    - 가장 유사한 template 선택 후 2D-2D matching을 통해 2D-3D correspondence 정의 후 PnP+RANSAC 적용
      - Rendered image의 2D-3D는 known information
    - 2D-2D matching loss를 사용하여, 6D pose estimation에서는 suboptimal함
  - MegaPose
    - 아래 두 과정을 통해 iterative pose update
      - 현재 추정된 자세로 3D 모델을 렌더링하여 2D 이미지 생성
      - Input image, render image를 6d parameter regression 네트워크에 입력
        - ✓Ground truth pose와 loss 계산
    - Pose regression과정은 네트워크가 학습한 시각적 패턴, 즉 데이터에 의존하며 projective geometry를 직접적으로 활용하지 않음 → unseen object에 대한 성능 저하로 이어짐

GenFlow[1)]

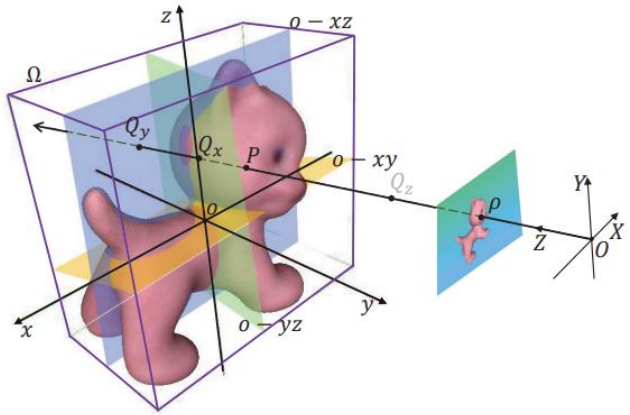
- 기존 모델의 optical flow 사용 방식
  - 3D shape 정보를 활용하지 않고 단순 2D 이미지 간 비교는 suboptimal 할 수 있음
  - Optical flow는 2D-2D correspondence 정보를 보유
    - Rendered image에 대한 2D-3D 대응은 알고 있으므로 이는 곧 2D-3D correspondence
    - 2D-3D correspondence를 안다면 geometry 기반 알고리즘을 통해 대수적으로 포즈 추정
  - Optical flow 사용 시 발생할 수 있는 문제점
    - 이미지의 밝기 변화량 즉, 겉보기 움직임을 추정하기 때문에 물리적으로 적합한 변화량이 아닐 수 있음

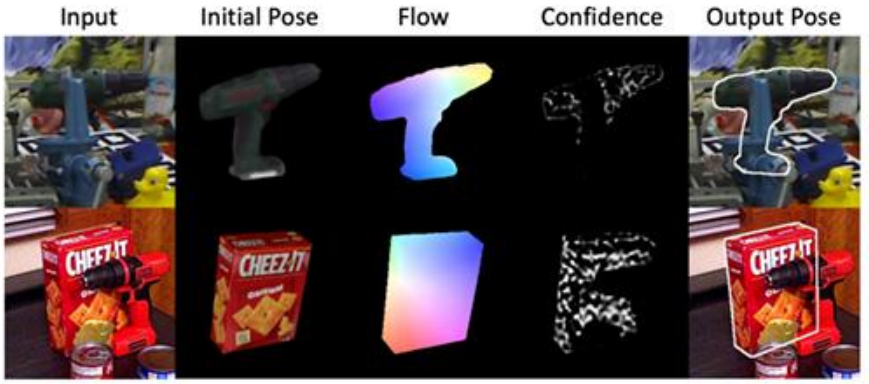




GenFlow[1)]

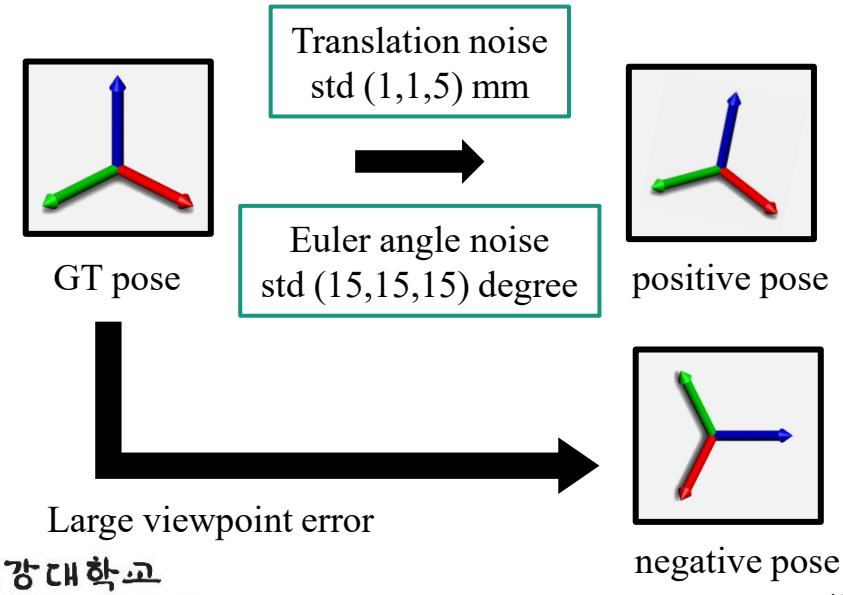
- 주요 특징
  - Optical flow 기반 pose refinement 방법론 고도화
    - Optical flow에 3D shape constraint 부여
    - Dense correspondence(flow)와 pose가 반복적으로 동시 업데이트
  - Differentiable PnP 알고리즘을 통한 end-to-end 학습
  - Disentangle loss / Confidence, pose sensitivity estimation / Cascaded structure

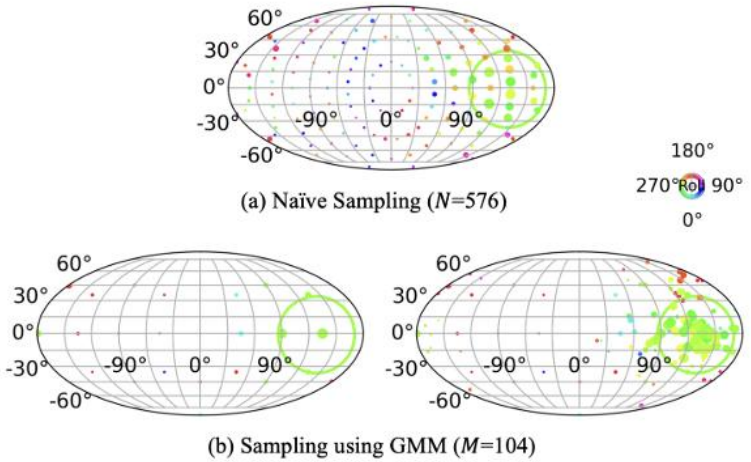




GenFlow[1)]

- Coarse pose estimation
  - 여러 각도에서 rendering된 574개의 template 중 하나를 선택
    - GT pose로부터 positive, negative pair를 생성하여 BCE loss를 통해 classifier를 학습
    - 해당 classifier는 pose 적합도에 대한 점수를 생성하는 scorer가 됨
  - 효율성을 위한 GMM (Gaussian Mixture Model) modeling 및 sampling
    - M개의 template만을 선정한 후, 상위 n개의 template으로 GMM modeling
  - Bounding box안에 대략적으로 위치하도록 translation을 대략 추정

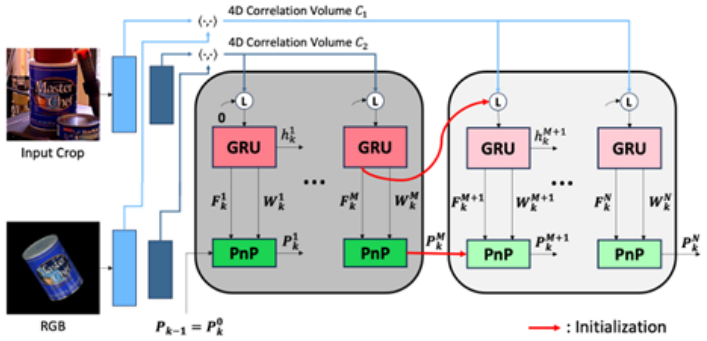


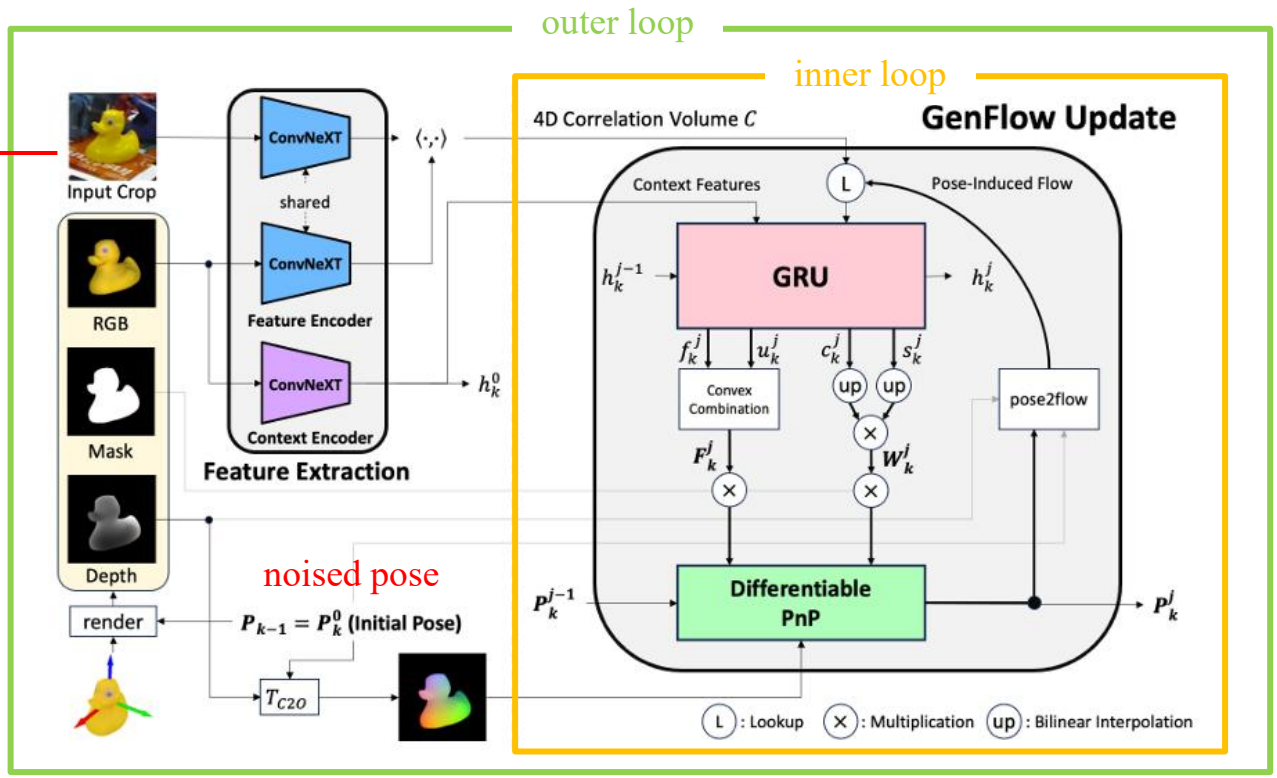


1) Moon, Sungphill, et al. "Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
2) Teed, Zachary, and Jia Deng. "Raft: Recurrent all-pairs field transforms for optical flow." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020.

GenFlow[1)]

- Pose refinement
  - GRU module from RAFT[2)]
  - Differentiable PnP
  - Pose2flow

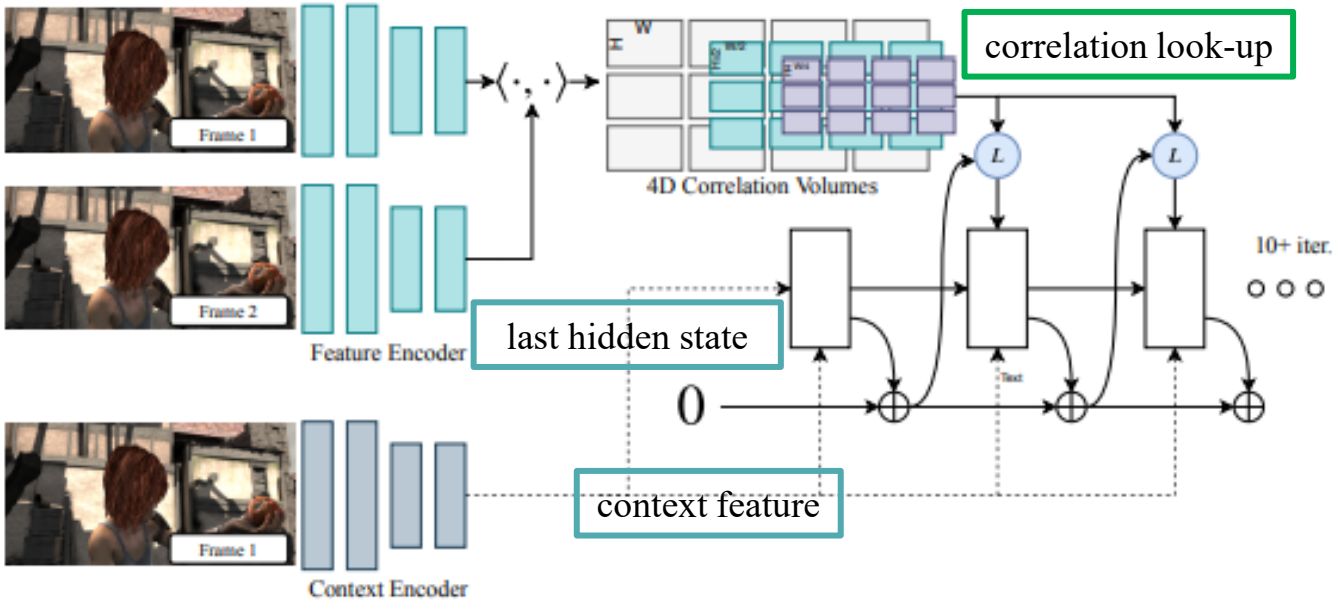




1) Moon, Sungphill, et al. "Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
2) Teed, Zachary, and Jia Deng. "Raft: Recurrent all-pairs field transforms for optical flow." Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer International Publishing, 2020.

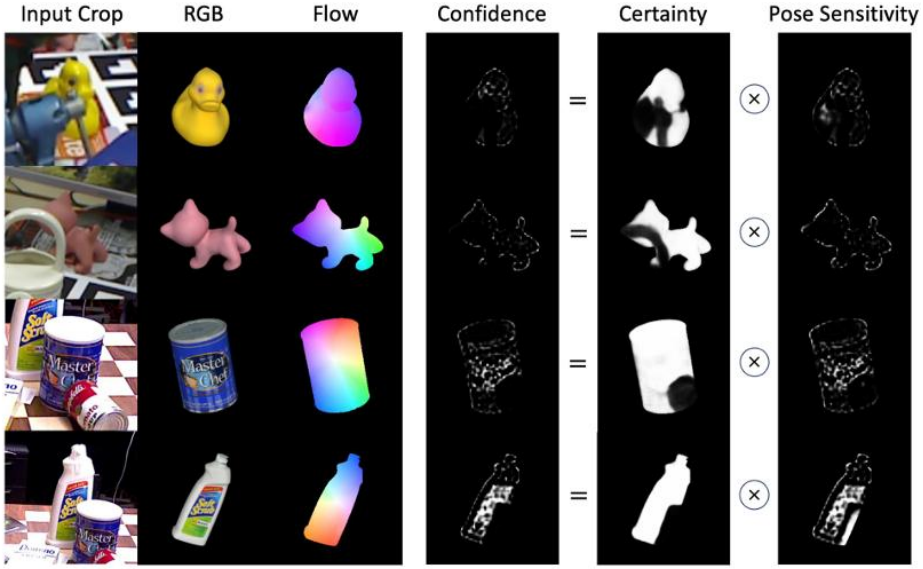
GenFlow[1)]

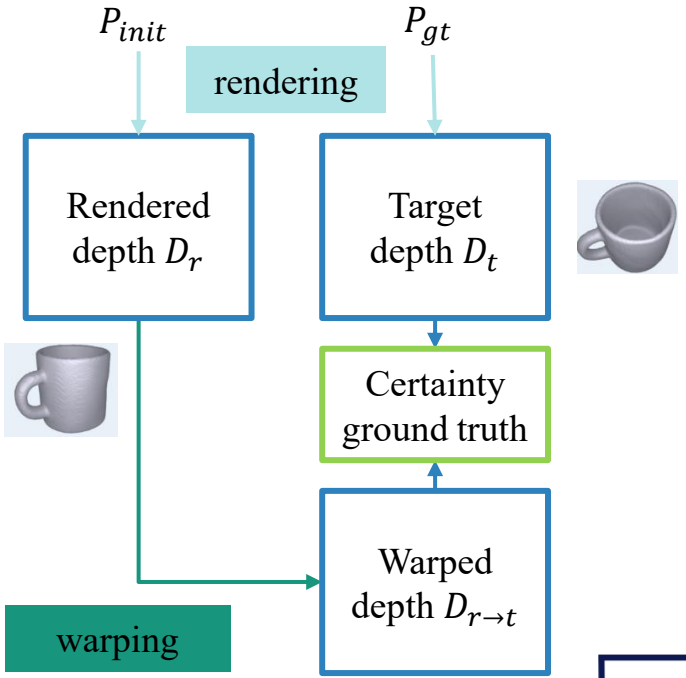
- Pose refinement
  - RAFT[2)]
    - Recurrent 방식의 GRU module을 사용하는 optical flow estimation model
    - 현재 추정된 optical flow에 대응하는 feature간의 correlation을 입력
      - 예상 대응점 주변의 이웃 값들을 가져와 종합하여 feature vector 생성
    - Up-sampling을 위해 convex combination 사용
      - GRU module이 출력하는 up-sampling mask로 3x3 grid에서 weighted sum 수행



GenFlow[1)]

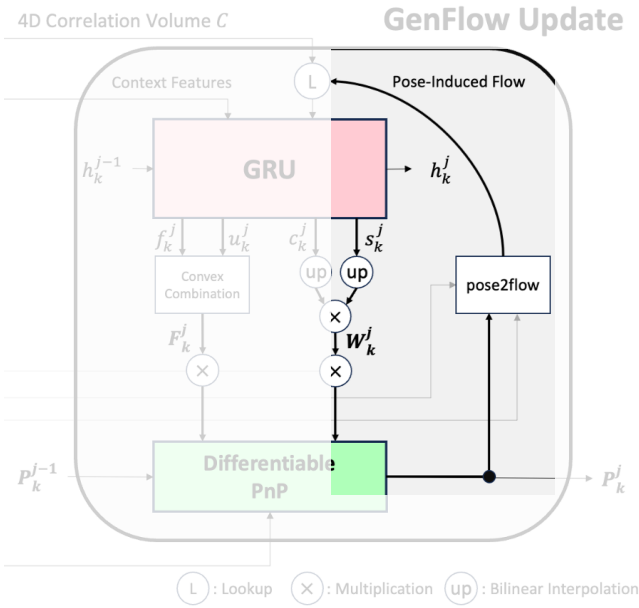
- Pose refinement
  - Confidence를 추정하기 위한 certainty, pose sensitivity head를 추가
    - Supervised by different losses - confidence factorization
  - Certainty c^j
    - Occlusion 발생여부를 추정하며 ground truth와의 cross-entropy loss인 L_{cert}를 통해서만 update
  - Pose sensitivity
    - Highlights the rich texture regions and extremities



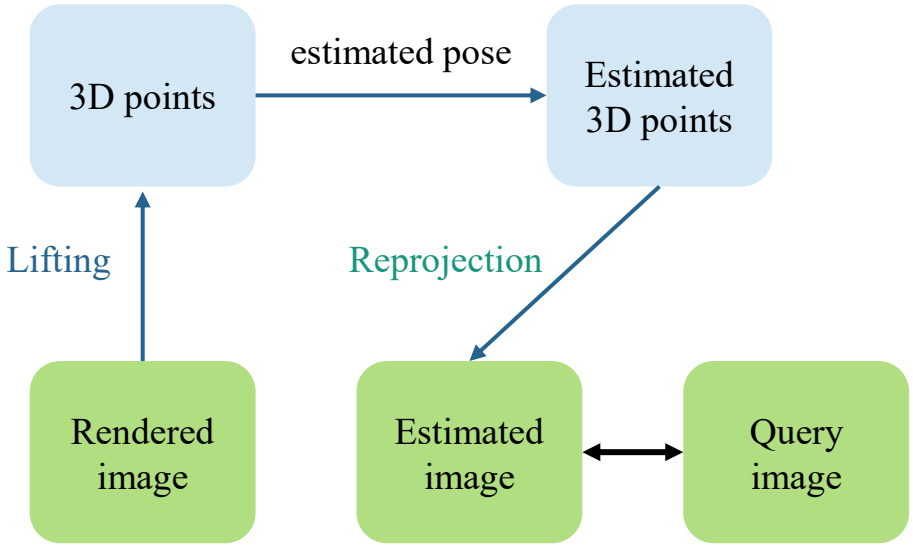


GenFlow[1)]

- Pose refinement
  - Pose2flow: pose induced flow
    - Optical flow 추정에 3D shape constraint를 주기 위한 핵심 부분
    - RAFT[2)] 구조는 현재 추정한 flow 정보를 look-up에 사용하여 다음 step에 반영
      - 추정된 2D optical flow가 아닌 pose 기반 flow를 전달하여 3D shape 정보 활용



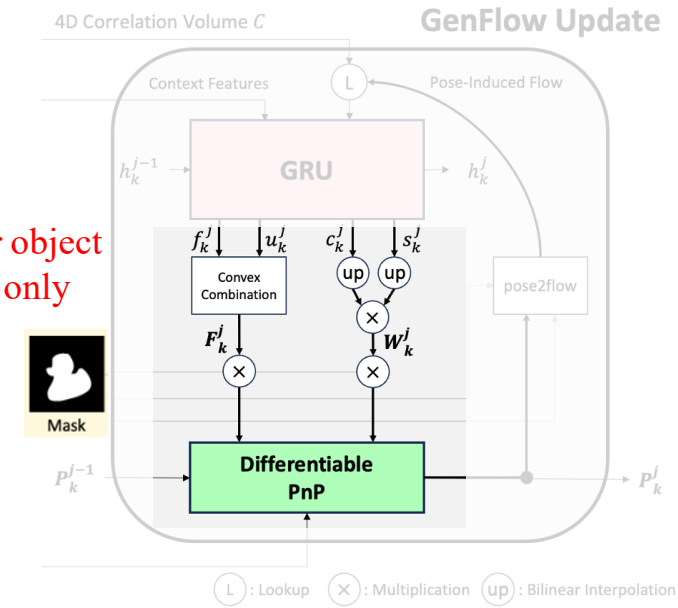
$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \mathbf{R}_k^{\mathrm{T}}(\mathbf{K}^{-1}\mathcal{D}_r(u,v)\begin{pmatrix} u \\ v \\ 1 \end{pmatrix} - \mathbf{t}_k)$$



pose-induced flow

GenFlow[1)]

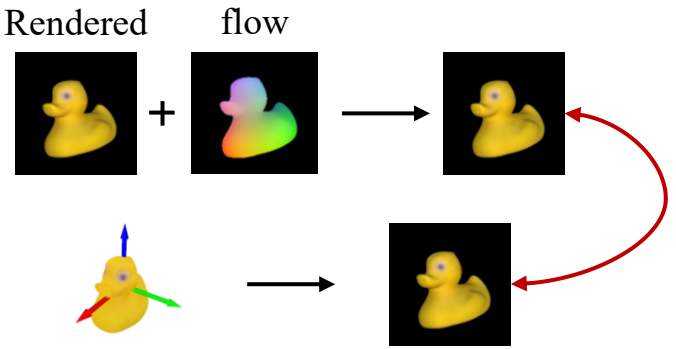
- Pose refinement
  - Pose estimation with PnP algorithm
    - Confidence weighted reprojection error minimize
    - Backpropagation되는 loss function과는 구분되어 연산 식 작성을 위해 존재
    - 3 LM step → detach pose P^* → Gauss-Newton method update P_k^j
      - LM 연산을 unrolling하여 backpropagation 시 numerical instability가 존재



Consider object region only

confidence

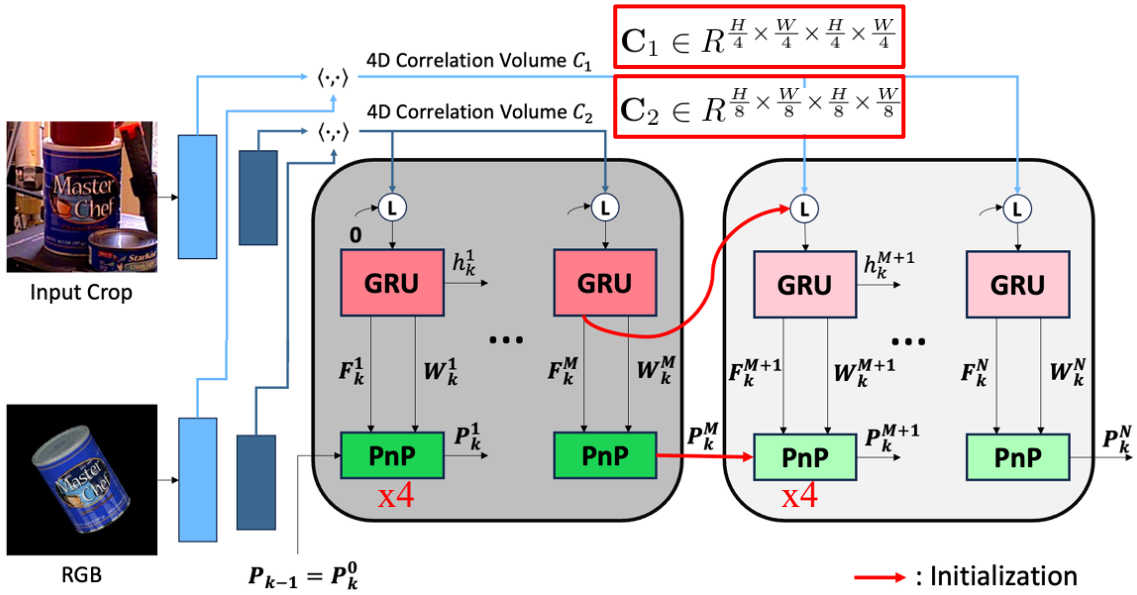
$$\underset{R,t}{\operatorname{argmin}} \frac{1}{2} \sum_u \sum_v \| \mathbf{W}^j(u,v) \times (\pi(R \begin{pmatrix} x \\ y \\ z \end{pmatrix} + t) - (\begin{pmatrix} u \\ v \end{pmatrix} + \mathbf{F}^j(u,v))) \|^2,$$



GenFlow[1)]

- Pose refinement
  - Cascaded architecture
    - Multi-scale을 고려하여 high-low level 순서로 optical flow 및 pose refine

Shape-constraint	$\mathcal{L}_{flow}, \mathcal{L}_{pose}$	64.9	81.8	75.1	57.0	70.1	69.8
Shape-constraint + Cascade	$\mathcal{L}_{flow}, \mathcal{L}_{pose}$	65.7	82.3	75.0	58.9	69.9	70.4
Shape-constraint + Confidence factorization	$\mathcal{L}_{flow}, \mathcal{L}_{cert}, \mathcal{L}_{pose}$	64.5	81.8	75.3	57.3	70.5	69.9
Shape-constraint + Confidence factorization + Cascade	$\mathcal{L}_{flow}, \mathcal{L}_{cert}, \mathcal{L}_{pose}$	65.9	82.0	76.1	59.5	69.8	**70.7**



GenFlow[1)]

- Pose refinement

$$\mathcal{L}_{pose} = \mathcal{D}([\mathbf{R}|[\mathbf{t}_x^\star, \mathbf{t}_y^\star, \mathbf{t}_z^\star]^{\mathrm{T}}], \mathbf{P}_{gt}) + \mathcal{D}([\mathbf{R}^\star|[\mathbf{t}_x, \mathbf{t}_y, \mathbf{t}_z^\star]^{\mathrm{T}}], \mathbf{P}_{gt}) + \mathcal{D}([\mathbf{R}^\star|[\mathbf{t}_x^\star, \mathbf{t}_y^\star, \mathbf{t}_z]^{\mathrm{T}}], \mathbf{P}_{gt})$$

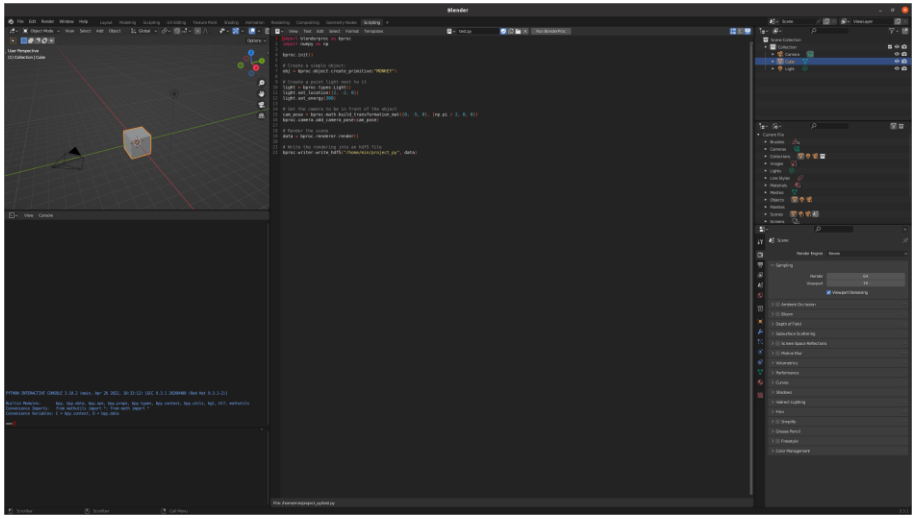
  - Loss function
    - Flow loss L_{flow}: Ground truth optical flow와의 L1 loss
    - Certainty loss L_{cert}: Binary cross entropy loss with certainty mask
    - Pose loss L_{pose} : Disentangled point matching loss
  - Additional Strategy1: Multi-hypothesis strategy
    - Coarse pose estimation에서 상위 n개 후보 선택 후 refine까지 수행
    - 최종 pose 후보들을 다시 coarse model에 넣어 가장 높은 점수의 pose 선택
  - Additional Strategy2: RGB-D input
    - Input depth map을 사용해 2D-3D를 3D-3D로 lift
    - 높은 confidence의 3D-3D correspondence만 남긴 후 RANSAC-Kabsch 알고리즘 적용

$$\mathcal{L} = \sum_{j=1}^{N} \gamma^{j-N} (\mathcal{L}_{flow}^{j} + \alpha \mathcal{L}_{cert}^{j} + \beta \mathcal{L}_{pose}^{j})$$

GenFlow[1)]

- Training detail
  - Dataset
    - 3D model: ShapeNet, Google Scanned Objects
    - Graphic tool: BlenderProc
      - Rendering, annotation 자동화 pipeline → 2 million RGB-D images generated

< ShapeNet >



< BlenderProc >

GenFlow[1)]

- BOP(Benchmark for 6D Object Pose Estimation) challenge
  - BOP 2023 new task: Unseen objects during training
  - 학습에 사용되지 않은 real world dataset으로 평가

BOP Benchmark for 6D Object Pose Estimation

- Evaluation metric
  - Average Recall, thresholding error from three error function
    - Visible Surface Discrepancy (VSD)
      - 보이는 부분에 대한 정렬 측정. Occlusion에 강건
    - Maximum Symmetry-Aware Surface Distance (MSSD)
      - 3D 모델 표면 위의 점 사이 거리 측정. Object symmetry 고려
    - Maximum Symmetry-Aware Projection Distance (MSPD)
      - 2D image 평면에서 오차 계산
    - 여러 가지 threshold를 적용하여 true/false 구분 및 averaging
      - Ex) object diameter의 5% ~ 50%

GenFlow[1)]

- Experiment result

	Method	Training	Average Recall (↑)					
			LM-O	T-LESS	TUD-L	IC-BIN	YCB-V	MEAN
1	No shape-constraint + RANSAC-PnP [39]	$\mathcal{L}_{flow}$	61.7	77.6	72.4	54.2	65.7	66.3
2	Shape-constraint	$\mathcal{L}_{flow}, \mathcal{L}_{pose}$	64.9	81.8	75.1	57.0	70.1	69.8
3	Shape-constraint + Cascade	$\mathcal{L}_{flow}, \mathcal{L}_{pose}$	65.7	82.3	75.0	58.9	69.9	70.4
4	Shape-constraint + Confidence factorization	$\mathcal{L}_{flow}, \mathcal{L}_{cert}, \mathcal{L}_{pose}$	64.5	81.8	75.3	57.3	70.5	69.9
5	Shape-constraint + Confidence factorization + Cascade	$\mathcal{L}_{flow}, \mathcal{L}_{cert}, \mathcal{L}_{pose}$	65.9	82.0	76.1	59.5	69.8	**70.7**

< Ablation >

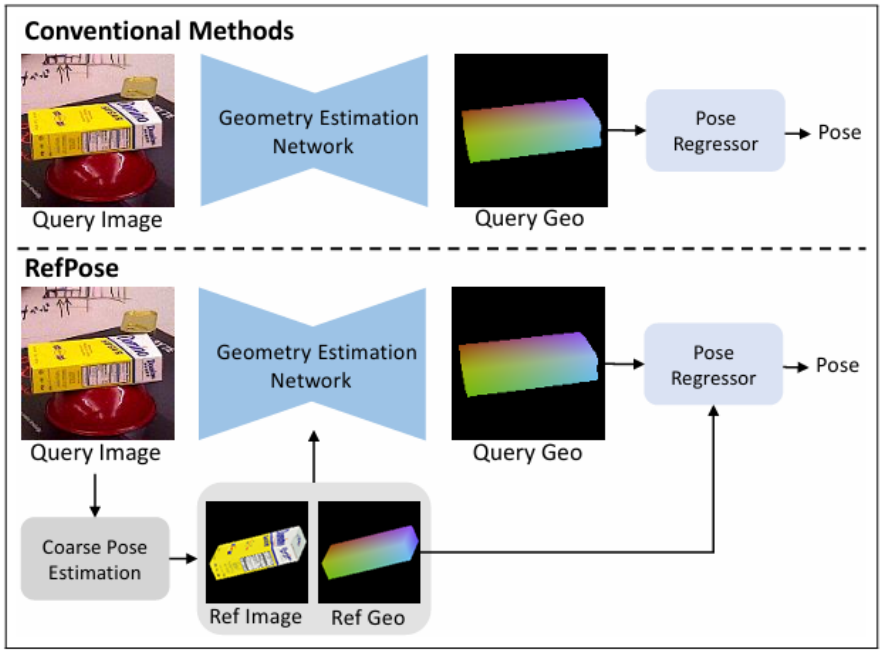
	2D Localization			Pose Initialization		Pose Refinement		Average Recall (↑)							
	RGB-D Input	Method	Novel Objects	Method	Novel Objects	Method	Novel Objects	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	MEAN
1	✗	Mask-RCNN [22, 38]	✗	CosyPose [38]	✗	CosyPose [38]	✗	63.3	64.0	68.5	58.3	21.6	65.6	57.4	57.0
2	✗	Mask-RCNN [22, 38]	✗	SurfEmb [21]	✗	BFGS	✗	66.3	73.5	71.5	58.8	41.3	79.1	64.7	65.0
3	✗	YOLOX [19, 44]	✗	ZebraPose [56]	✗	-	✗	**72.9**	81.1	75.6	59.2	**50.4**	**92.1**	**72.9**	**72.0**
4	✗	YOLOX [19, 44]	✗	MegaPose [37]	✓	MegaPose+MH [37]	✓	64.8	78.1	74.1	56.9	42.2	86.3	70.2	67.5
5	✗	YOLOX [19, 44]	✗	Ours	✓	Ours+MH	✓	68.3	**82.8**	**77.8**	**59.6**	50.1	89.7	70.8	71.3
6	✓	YOLOX [19, 44]	✗	WDR-Pose [30]	✗	PFA [31]+Kabsch	✗	**79.2**	**84.9**	**96.3**	**70.6**	52.6	86.7	**89.9**	**80.0**
7	✓	YOLOX [19, 44]	✗	MegaPose [37]	✓	MegaPose+MH [37]+Teaserpp [67]	✓	70.4	71.8	91.6	59.2	55.3	87.2	85.5	74.4
8	✓	YOLOX [19, 44]	✗	Ours	✓	Ours+MH+Kabsch	✓	74.2	78.3	92.8	64.9	**65.2**	**92.0**	88.3	79.4
9	✗	OSOP [54]	✓	OSOP [54]	✓	OSOP+PnP+MH [54]	✓	31.2	-	-	-	-	49.2	33.2	-
10	✗	CNOS-det. [50]	✓	MegaPose [37]	✓	MegaPose+MH [37]	✓	56.0	50.8	68.7	41.9	34.6	70.6	62.0	54.9
11	✗	CNOS-det. [50]	✓	Ours	✓	Ours+MH	✓	**57.5**	**53.0**	**69.1**	**45.6**	**40.8**	**74.5**	**63.9**	**57.8**
12	✓	OSOP [54]	✓	OSOP [54]	✓	OSOP+Kabsch+MH [54]+ICP	✓	48.2	-	-	-	-	60.5	57.2	-
13	✓	CNOS-seg. [50]	✓	ZeroPose [5]	✓	MegaPose+MH [37]	✓	53.8	40.0	83.5	39.2	52.1	65.3	65.3	57.0
14	✓	CNOS-det. [50]	✓	MegaPose [37]	✓	MegaPose+MH [37]+Teaserpp [67]	✓	62.6	48.7	85.1	46.7	46.8	73.0	76.4	62.8
15	✓	CNOS-det. [50]	✓	Ours	✓	Ours+Kabsch+MH	✓	**62.9**	**51.7**	**85.8**	**53.3**	**55.9**	**78.2**	**82.5**	**67.2**

< BOP challenge >

- RefPose: Leveraging Reference Geometric Correspondences for Accurate 6D Pose Estimation of Unseen Objects (CVPR 2025)

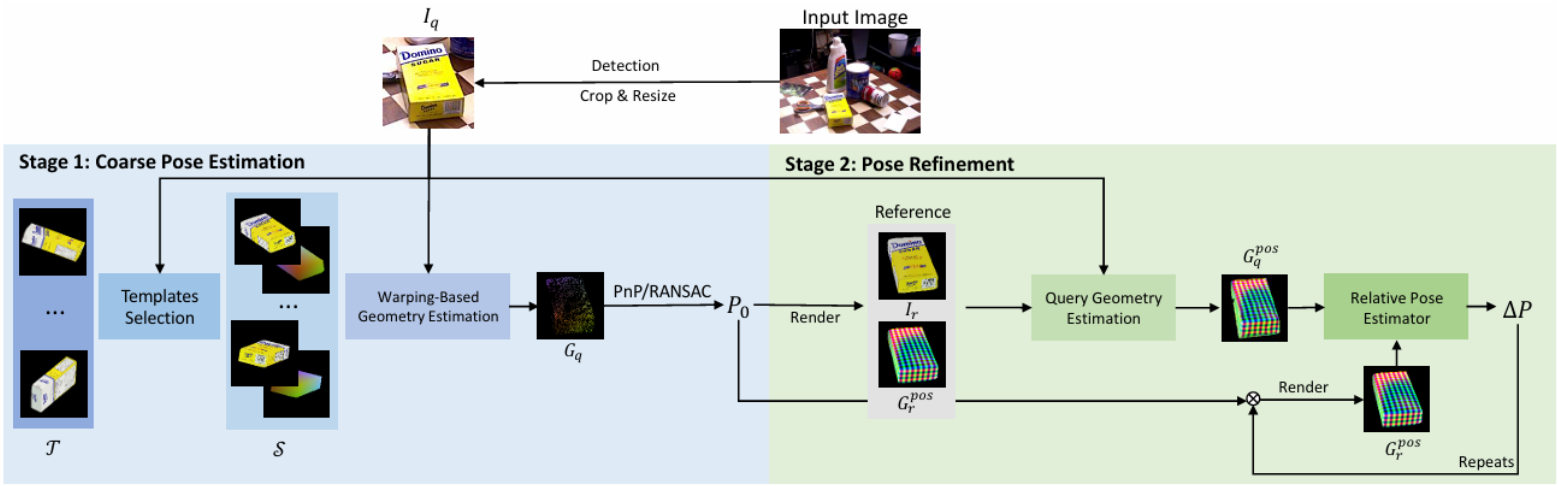
RefPose[1)]

- Unseen object에 대한 일반화 성능 향상
  - Query image로부터 바로 geometric correspondence를 추정
    - Shape prior에 대한 의존도가 높아 unseen object에 대한 일반화 성능이 떨어짐
  - Reference image를 같이 입력하는 방식의 네트워크 (Render-and-compare)
    - Pre-rendered template에 대한 문제제기
    - Template 자체의 성능 부족 가능성 및 3D shape 정보 활용의 부재



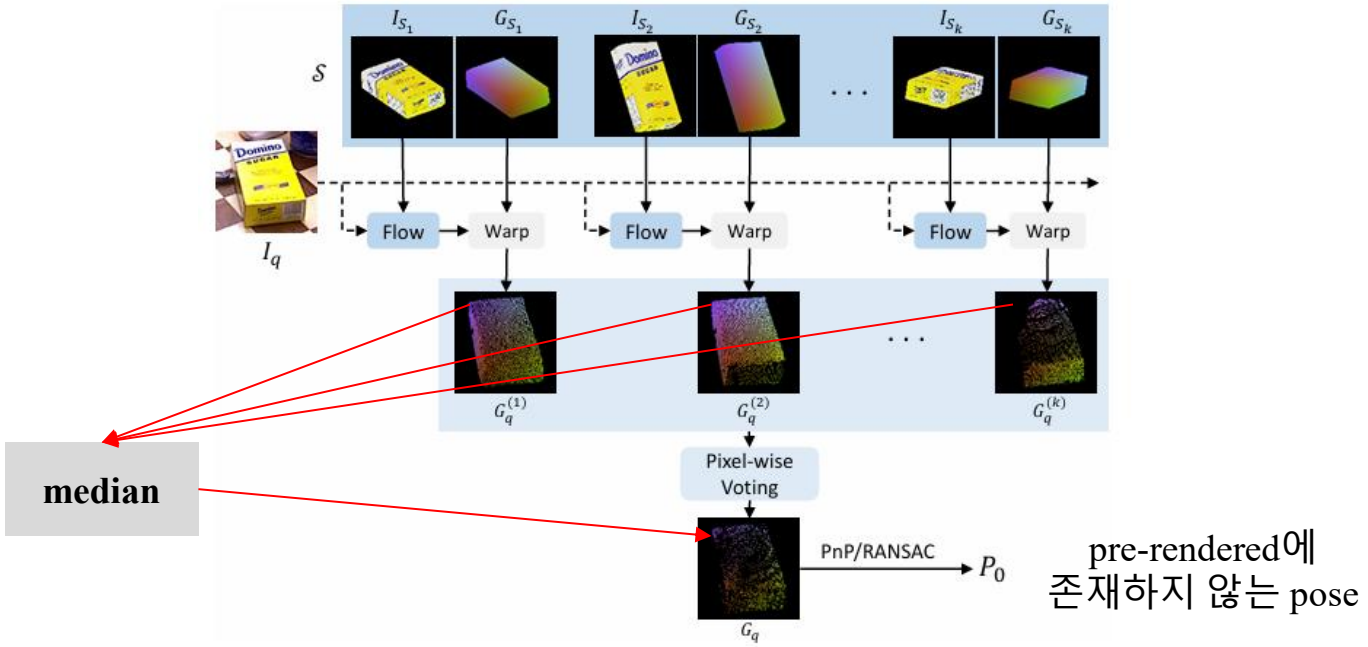
RefPose[1)]

- Unseen object에 대한 일반화 성능 향상
  - Optical flow 활용: Coarse pose estimation 강화
    - Optical flow를 통해 pre-rendered template에 추가 processing을 가해 reference image 생성
      - Better initial pose에서부터 refinement 가능
    - Reference image의 geometry(dense coordinate map)을 입력하여 guidance 강화
  - Geometry 기반 pose refinement



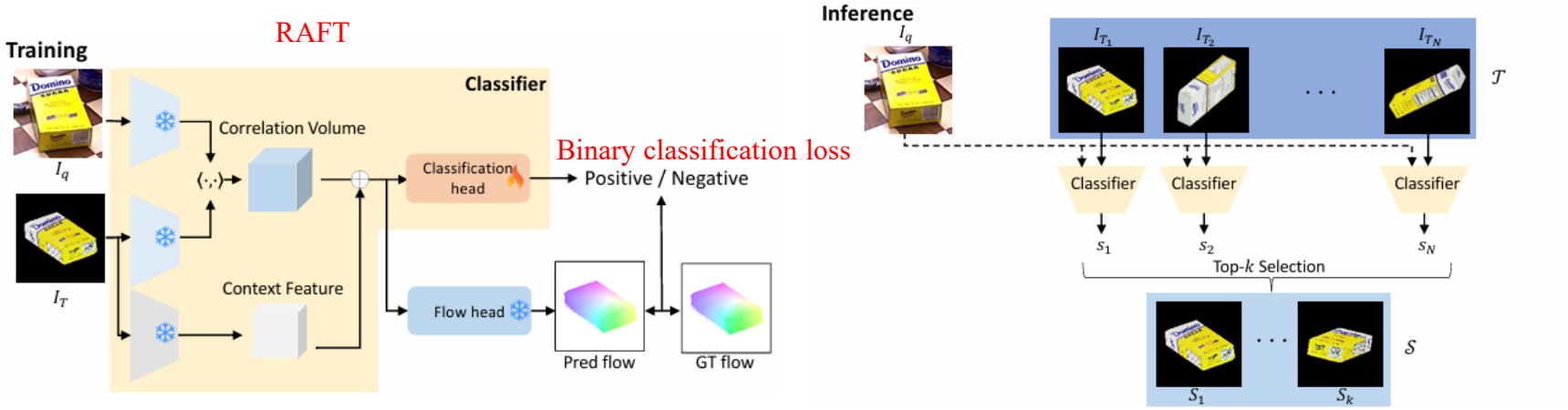
RefPose[1)]

- Coarse pose estimation
  - Classifier를 사용해 template scoring → top k template 선정
  - Query image와의 flow를 계산하여 template의 dense coordinate map을 warping
    - [2D coordinate, 3D coordinate] → [warped 2D coordinate, 3D coordinate]
  - Pixel-wise voting을 통해 최종 geometry 선정 및 initial pose 추정



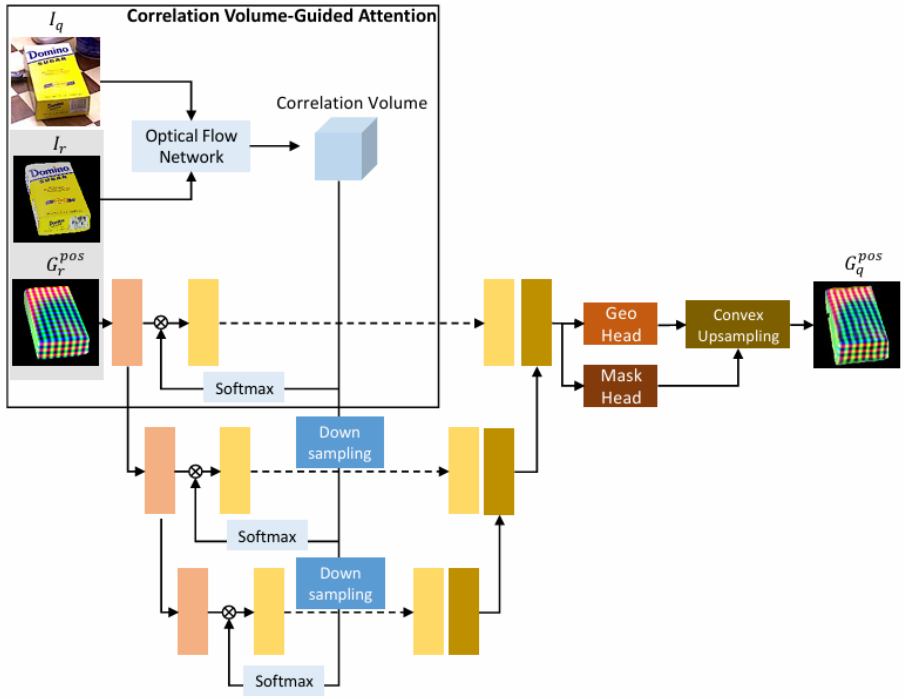
RefPose[1)]

- Coarse pose estimation
  - Query와 template간 flow가 얼마나 정확하게 예측될 수 있는 가를 기준으로 평가
    - Query와 가까운 pose를 가진 template 이라는 조건은 간접적으로 반영됨
  - Pred flow와 GT flow간 차이에 따라 positive, negative label 부여
    - Predicted flow와 ground truth flow간의 오차를 supervision으로 활용
    - BCE loss로 가벼운 CNN 구조의 classification head를 학습시킴

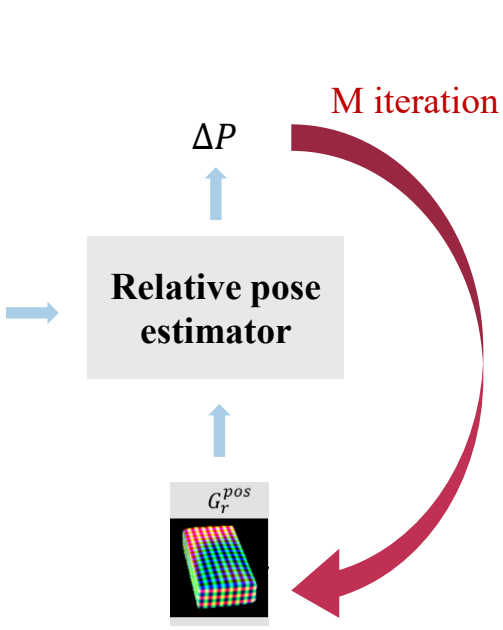


RefPose[1)]

- Pose refinement
  - Initial pose에 해당하는 geometry로부터 query geometry estimation
    - Dense coordinate map을 deep network에서 효과적으로 사용하기 위한 positional encoding
      - 3차원 좌표를 다양한 주파수의 sinusoid 에 통과시켜 더 고차원의 벡터로 encoding
  - Relative pose estimator를 통해 query geometry의 pose 추정



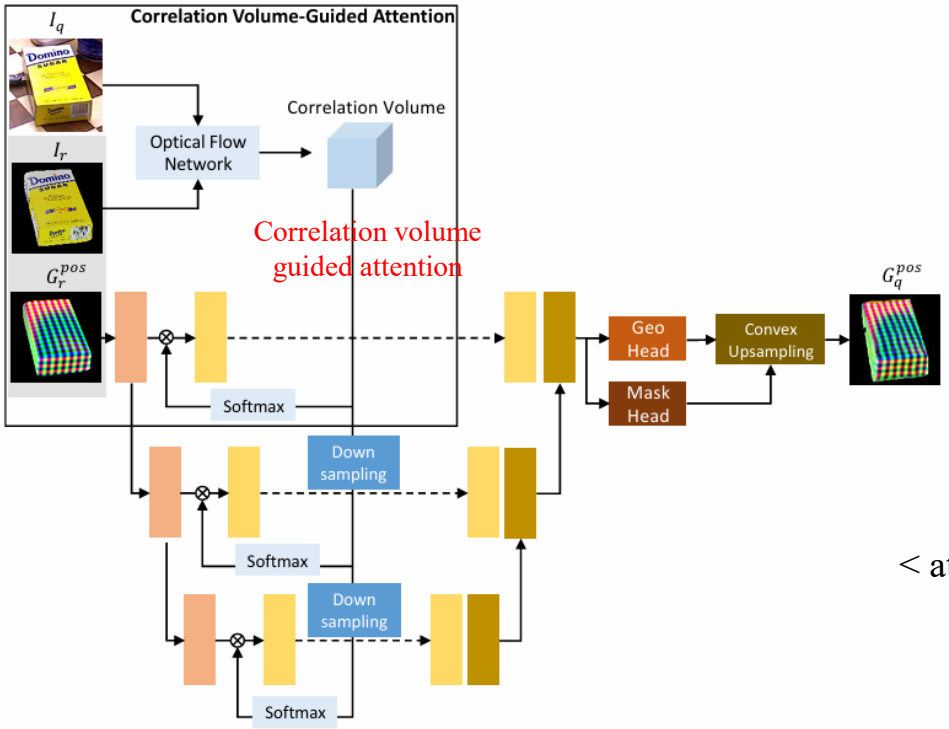
< Geometry estimation >

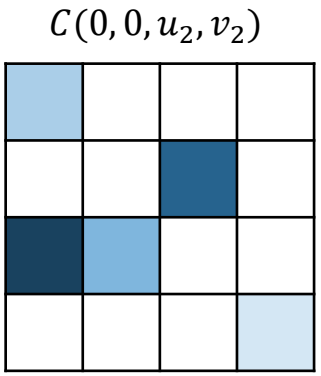


< Pose alignment >

RefPose[1)]

- Pose refinement
  - Correlation volume-guided attention
    - Q: query image feature / K: reference image feature / V: G_r^{pos}
  - Up-sampling mask를 통해 geometry up-sampling하여 최종 geometry 예측





< attention weight for G_r^{pos} (0,0) processing >

RefPose[1)]

- Pose refinement
  - Loss design

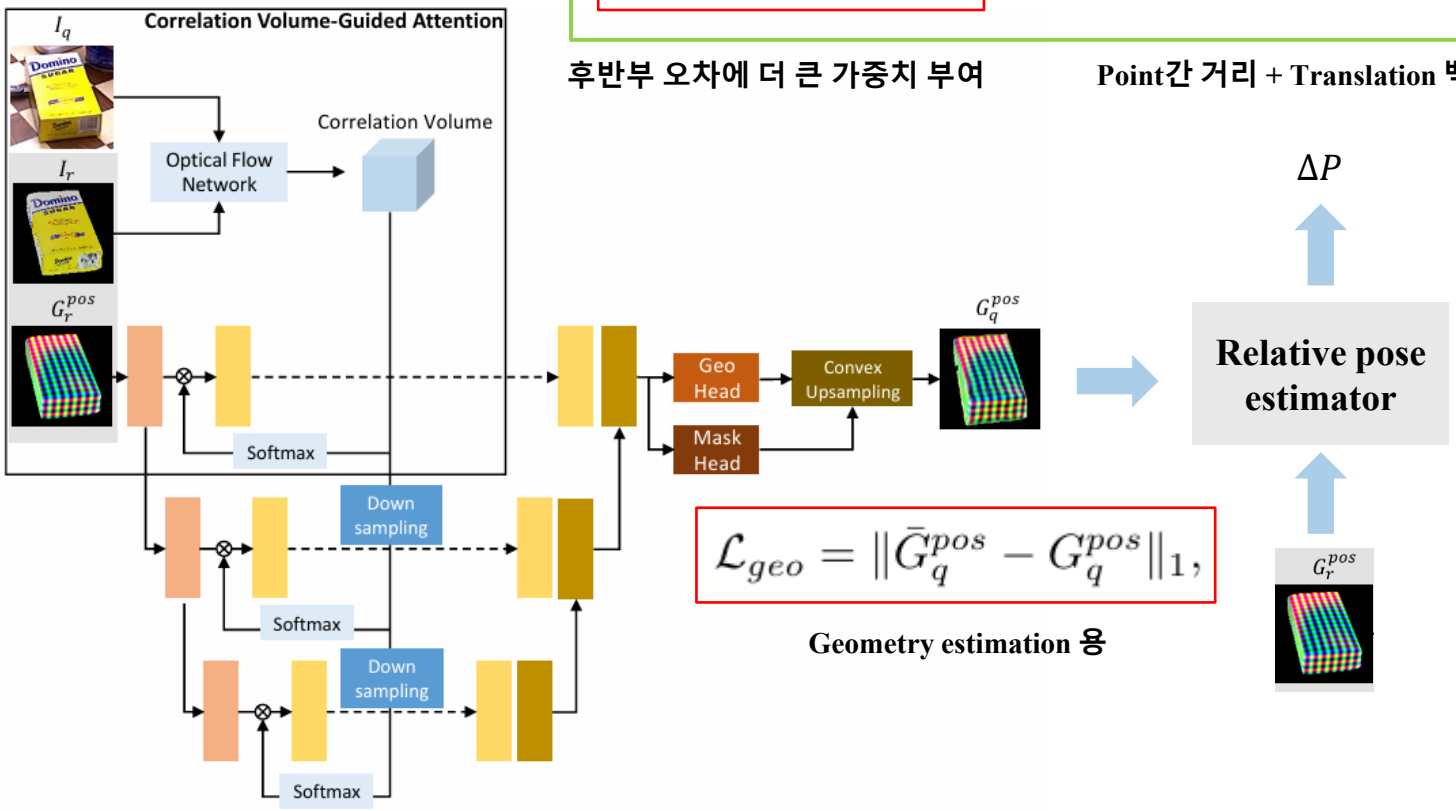
$$P + \Delta P \iff \text{Ground truth pose}$$

$$\mathcal{L}_{seq} = \sum_{m=1}^{M} \gamma^{M-m} \mathcal{L}_{pose}^{(m)}$$

$$\mathcal{L}_{pose} = \|\bar{\mathcal{G}} - \mathcal{G}\|_2 + \|\|\bar{t}\|_2 - \|t\|_2\|_1$$

후반부 오차에 더 큰 가중치 부여

Point간 거리 + Translation 벡터 간 거리



$$\mathcal{L}_{geo} = \|\bar{G}_q^{pos} - G_q^{pos}\|_1,$$

Geometry estimation 용

RefPose[1)]

- Experiment result

Method	Refinement	YCB-V	Mean	Run-time
OSOP [41]	-	29.6	-	-
ZS6D [1]	-	32.4	-	-
MegaPose [15]	-	28.1	20.8	15.5s
GenFlow [28]	-	27.7	23.	3.8s
GigaPose [30]	-	27.8	26.8	**0.4s**
FoundPose [32]	-	45.2	37.2	1.7s
RefPose (Ours)	-	**50.0**	**38.1**	3.1s
MegaPose [15]	MegaPose	60.1	50.9	17.0s
MegaPose [15]	MegaPose, MH	62.1	54.7	21.9s
MegaPose [15]	RefPose (Ours)	65.3	57.2	16.4s
GenFlow [28]	GenFlow, MH	63.3	57.0	20.8s
GigaPose [30]	MegaPose	63.2	54.7	**2.3s**
GigaPose [30]	GenFlow, MH	65.2	60.5	10.6s
FoundPose [32]	MegaPose, MH + Featuremetric	69.0	59.6	20.5s
RefPose (Ours)	MegaPose	63.7	56.3	4.6s
RefPose (Ours)	RefPose (Ours)	**72.7**	**61.4**	3.9s

Conclusion

- GenFlow
  - Recurrent flow를 통한 점진적 정제
  - Key contribution
    - GRU 기반 재귀적 구조
    - Cascaded refinement
    - GMM 기반 초기 자세 샘플링
- RefPose
  - Geometric correspondence를 직접 예측
  - Key contribution
    - Optical flow 기반 template 선택
    - Iterative reference comparison
    - Correlation volume-guided attention